Financial time series

François Roueff

January 29, 2024

Contents

Foreword

These lecture notes serve as a material for the course on financial time series of the M2 SFA of Institut Polytechnique de Paris. Time series analysis is widespread in various applications ranging from engineering sciences to social sciences such as econometrics, climatology, hydrology, signal processing, Internet metrology, and so on. For this reason and because many theoretical problems and practical issues remain unsolved, it has become an important field of study in the domain of statistics and probability. Here, while providing the foundations in a quite general fashion, we focus on models and approaches that have been popular for studying econometric or financial time series, see Tsay [2005] for a classical introduction with many practical examples.

Part I can essentially be seen as a summary of the prerequisites for this course. It should nevertheless provide a solid introduction to the basic principles of weakly stationary processes, ARMA processes and statistical inference for time series. Essential references for students interested in these topics are Brockwell and Davis [1991] and Shumway and Stoffer [2011]. In this first part, we mainly consider linear models. We start by setting the general framework of stochastic modelling in Chapter 1. We then focus on second order properties in Chapter 2 and linear models in Chapter 3, with a detailed description of the ARMA model. Finally statistical inference for analyzing temporally dependent data are introduced in Chapter 4.

Part II constitutes the main content of these lecture notes. It basically provides the theoretical background behind the classical examples discussed in [Tsay, 2005, Chapters 2,3,8,11]. In Chapter 5, the most widespread models such as GARCH processes for univariate financial time series are introduced and briefly discussed. In Chapter 6, we explain how the tools introduced for univariate (or real valued) time series can without much effort be extended to the case where the time series are valued in an abstract Hilbert space. Although we are more specifically interested in the finite dimensional case, that is, to *multivariate time series* in the following but it is interesting to understand that many concepts are easy to adapt to the infinite dimensional case, since this setting is now widespread to deal with *functional data*, see Horváth and Kokoszka [2012]. In Chapter 7, we introduce the framework of *linear state space models*, or *dynamic linear models*. Numerical algorithms for forecasting will be derived in this context. Such methods apply to a large class of multivariate models and can be used also for statistical inference. The models covered by this chapter extend the ARMA models in a natural way, albeit in a multivariate setting. An important problem arising in financial and econometric time series, as in any application where forecasting is at stake, is to investigate the stability in the data, or, to put it in statistical way, is the data stationary? Chapter 8 provides the natural extension of ARMA models which allows to give a meaningful statistical answer about how stationary the time series is. In the case of multivariate time series, the same questions turn out to be much more difficult to answer. In Chapter 9, we investigate some answers applying to this case. A vary complete view on this topic can be found in Lütkepohl [2005].

Finally, Appendix A, Appendix B and Appendix C provide brief accounts of the essential definitions and results on Hilbert spaces, conditional distributions and on the convergence of random variables.

Most of the numerical experiments which illustrate these notes have been performed using the software R (see R software) and sometimes required the packages astsa and fGarch.

Notation and conventions

Vectors of $\mathbb{C}^d$ are identified to $d \times 1$ matrices.

The component-wise conjugate of $x \in \mathbb{C}^d$ is denoted by $\bar{x}$.

The Hermitian norm of $x \in \mathbb{C}^d$ is denoted by $|x|$.

The identity operator and matrix are denoted by $\mathbf{1}$ (or $\mathbf{1}_d$ to precise its dimension $d \times d$)

The transpose of matrix A is denoted by A^T.

The conjugate transpose of matrix A is denoted by A^H.

The set $\mathbb{T}$ is the quotient space $\mathbb{R}/(2\pi\mathbb{Z})$ (or any interval congruent to $[0, 2\pi)$).

The variance of the random variable X is denoted by $\mathrm{Var}\,(X)$.

The variance-covariance matrix of the random vector $\mathbf{X}$ is denoted by $\mathrm{Cov}(\mathbf{X})$.

The covariance matrix between the random vectors $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})$.

The Gaussian distribution with mean μ and covariance Q is denoted by $\mathcal{N}(\mu, Q)$.

$X \sim P$ means that the random variable X has distribution P

For a r.v. X on $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{P}^X$ denotes the probability distribution of X, $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$.

$(X_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$ means that $(X_t)_{t\in\mathbb{Z}}$ is a weak white noise with variance σ^2

$(X_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0, \sigma^2)$ means that $(X_t)_{t\in\mathbb{Z}}$ is a strong white noise with (finite) variance σ^2

$(X_t)_{t\in T} \overset{\text{iid}}{\sim} P$ means that $(X_t)_{t\in T}$ are independent variables with common distribution P.

Given $f : \mathbb{R}^p \to \mathbb{R}^q$ differentiable, $\partial f : \mathbb{R}^p \to \mathbb{R}^{p\times q}$ is the gradient of each component of f stacked columnwise.

Pursuing with the former example, if f is twice differentiable, $\partial\partial^T f : \mathbb{R}^p \to \times\mathbb{R}^p \times \mathbb{R}^{p\times q}$ are the Hessian matrices conveniently stacked depending of the context.

X_n converges a.s., in probability or weakly to X is denoted by $X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{P} X$ or $X_n \Longrightarrow X$, respectively.

The finite distributions of X_n converge weakly to that of X is denoted by $X_n \overset{\text{fidi}}{\Longrightarrow} X$.

Functions and measures spaces symbols

The set of all measurable functions defined on $(\mathsf{X}, \mathcal{X})$ and valued in $(\bar{\mathbb{R}}_+, \mathcal{B}(\bar{\mathbb{R}}_+))$ is denoted by $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$

The set of all bounded continuous functions defined on the metric space (X, d) and valued in $\mathbb{R}$ is denoted by $\mathrm{C}_b(\mathsf{X}, d)$

The set of all continuous linear functions defined from E to F is denoted by $\mathcal{L}_b(E, F)$

The set of all Lipschitz functions and valued in $\mathbb{R}$ defined on the metric space (X, d) is denoted by $\mathrm{Lip}(\mathsf{X}, d)$

The set of all bounded and Lipschitz functions defined on the metric space (X, d) and valued in $\mathbb{R}$ is denoted by $\mathrm{Lip}_b(\mathsf{X}, d)$

The set of all non-negative σ-finite measures defined on $(\mathsf{X}, \mathcal{X})$ is denoted by $\mathbb{M}_+(\mathsf{X}, \mathcal{X})$

The set of all probability measures defined on $(\mathsf{X}, \mathcal{X})$ is denoted by $\mathbb{M}_1(\mathsf{X}, \mathcal{X})$

In all above definitions, when no ambiguity arises, we will often omit the second arguments $\mathcal{X}$ or d *e.g.* simply writing $\mathrm{C}_b(\mathsf{X})$ or $\mathbb{M}_1(\mathsf{X})$.

Part I

Reminders on time series

Chapter 1

Random processes

In this chapter, we introduce the basic foundations for stochastic modelling of time series such as random processes, stationary processes, Gaussian processes and finite distributions. We also provide some basic examples of real life time series.

1.1 Introduction

A time series is a sequence of observations x_t, each of them recorded at a time t. The time index can be discrete, in which case we will take $t \in \mathbb{N}$ or $\mathbb{Z}$ or can be continuous, $t \in \mathbb{R}$, $\mathbb{R}_+$ or $[0, 1]$... Time series are encountered in various domains of application such as medical measurements, telecommunications, ecological data and econometrics. In some of these applications, spatial indexing of the data may also be of interest. Although we shall not consider this case in general, many aspects of the theory and tools introduced here can be adapted to spatial data.

In this course, we consider the observations as the realized values of a random process $(X_t)_{t\in T}$ as defined in Section 1.2. In other words, we will use a *stochastic modeling* approach of the data. Here are some examples which illustrate the various situations in which stochastic modelling of time series are of primary interest.

Example 1.1.1 (Heartbeats)**.** *Figure 1.1 displays the heart rate of a resting person over a period of* 900 *seconds. This rate is defined as the number of heartbeats per unit of time. Here the unit is the minute and is evaluated every* 0.5 *seconds.*

Example 1.1.2 (Internet traffic)**.** *Figure 1.2 displays the inter-arrival times of TCP packets, expressed in seconds, on the main link of Lawrence Livermore laboratory. This trace is obtained from a 2 hours record of the traffic going through this link. Over this period around 1.3 millions of packets have been recorded. Many traces are available on* The Internet Traffic Archive, *http://ita.ee.lbl.gov/*.

Example 1.1.3 (Speech audio data)**.** *Figure 1.3 displays a speech audio signal with a sampling frequency equal to 8000 Hz. This signal is a record of the unvoiced fricative phoneme* sh *(as in* sh*arp).*

Example 1.1.4 (Meteorological data)**.** *Figure 1.4 displays the daily record of the wind speed at the Kilkenny meteorological station.*

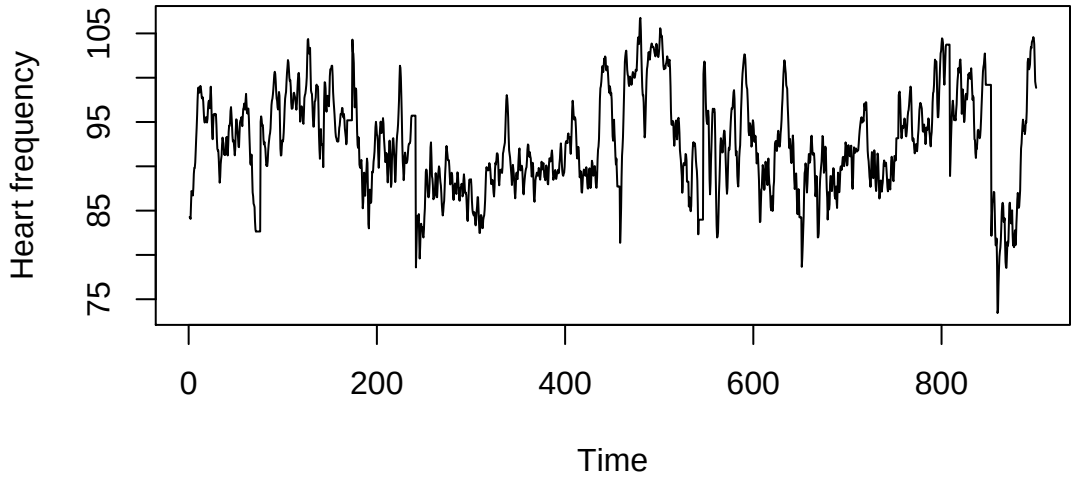

Figure 1.1: *Heartbeats: time evolution of the heart rate.*

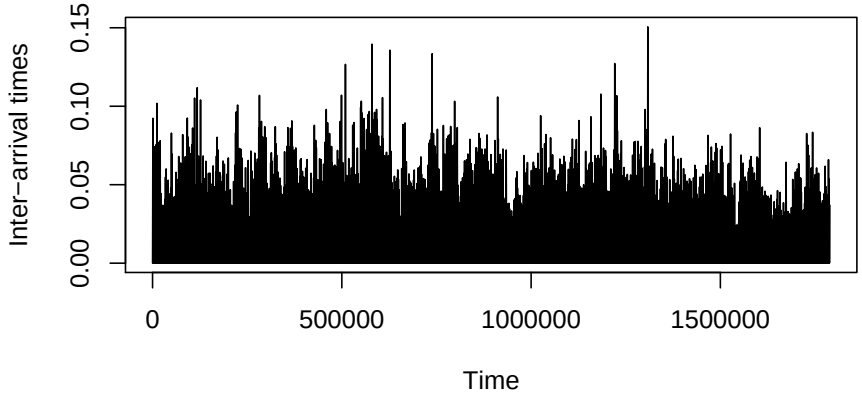

Figure 1.2: *Internet traffic trace : inter-arrival times of TCP packets.*

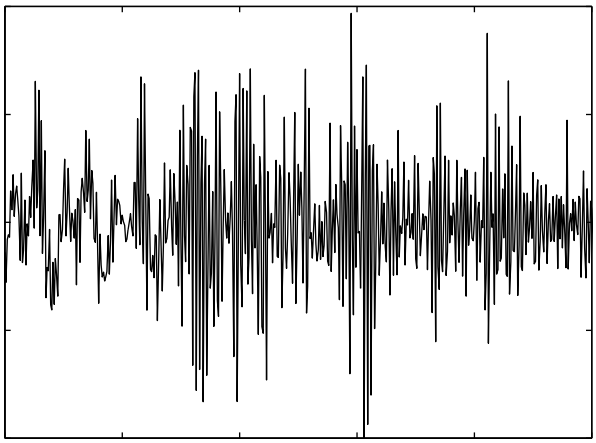

Figure 1.3: *A record of the unvoiced fricative phoneme* sh.

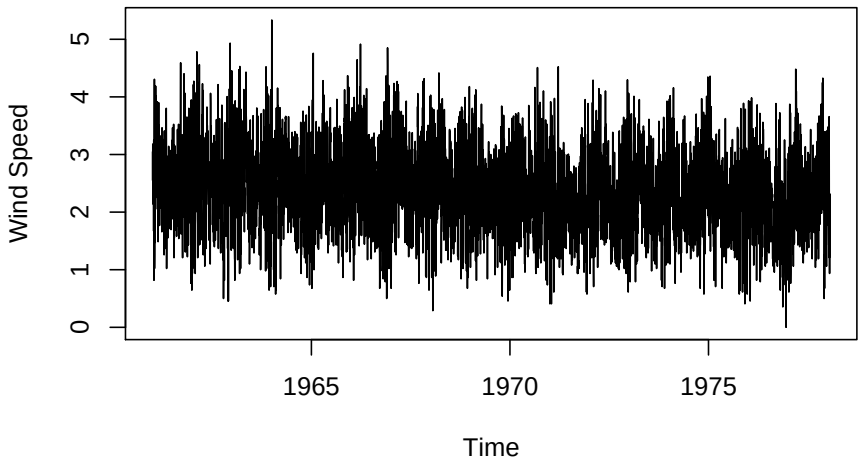

Figure 1.4: *Daily record of the wind speed at Kilkenny (Ireland).*

Example 1.1.5 (Financial index)**.** *Figure 1.5 displays the daily open value of the Standard and Poor 500 index. This index is computed as a weighted average of the stock prices of 500 companies traded at the New York Stock Exchange (NYSE) or NASDAQ. It is a widely used benchmark index which provides a good summary of the U.S. economy.*

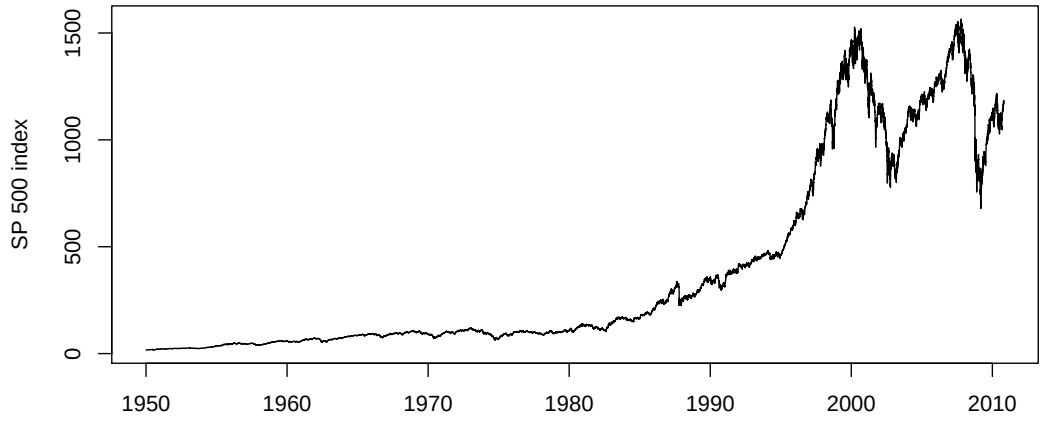

Figure 1.5: SP-500 stock index time series

1.2 Random processes

1.2.1 Definitions

In this section we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an index set T and a measurable space $(\mathsf{X}, \mathcal{X})$, called the *observation space*.

Definition 1.2.1 (Random process)**.** *A random process defined on $(\Omega, \mathcal{F}, \mathbb{P})$, indexed on T and valued in $(\mathsf{X}, \mathcal{X})$ is a collection $(X_t)_{t\in T}$ of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking there values in $(\mathsf{X}, \mathcal{X})$.*

The index t can for instance correspond to a time index, in which case $(X_t)_{t\in T}$ is a time series. When moreover $T = \mathbb{Z}$ or $\mathbb{N}$, we say that it is a *discrete time* process and when $T = \mathbb{R}$ or $\mathbb{R}_+$, it is a *continuous time* process. In the following, we shall mainly focus on discrete time processes with $T = \mathbb{Z}$. Concerning the space $(\mathsf{X}, \mathcal{X})$, we shall usually consider $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-field of $\mathbb{R}$), in which case we have a *real-valued process*, or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, in which case we have a *vector-valued process*, and in particular $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$, in which case we have a *complex-valued process*.

It is important to note that a random process can be seen as an application $X : \Omega \times T \to \mathsf{X}$, $(\omega, t) \mapsto X_t(\omega)$ such that, for each index $t \in T$, the function $\omega \mapsto X_t(\omega)$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathsf{X}, \mathcal{X})$.

Definition 1.2.2 (Path)**.** *For each $\omega \in \Omega$, the $T \to \mathsf{X}$ application $t \mapsto X_t(\omega)$ is called the* path *associated to the experiment ω.*

When $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$, it can be useful to associate a *filtration* to the process.

Definition 1.2.3 (Filtration)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$.*

(i) A filtration of a measurable space $(\Omega, \mathcal{F})$ is an increasing sequence $(\mathcal{F}_t)_{t\in T}$ of sub-σ-fields of $\mathcal{F}$. A filtered probability space $(\Omega,\ \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbb{P})$ is a probability space endowed with a filtration.

(ii) A random process $(X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be adapted to the filtration $(\mathcal{F}_t)_{t\in T}$ if for each $t \in T$, X_t is $\mathcal{F}_t$-measurable.

We will also directly say that $((X_t, \mathcal{F}_t))_{t\in T}$ is an adapted process to indicate that the process $(X_t)_{t\in T}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in T}$. The σ-field $\mathcal{F}_t$ can be thought of as the information available up to time t. Requiring the process to be adapted means that X_t can be computed using this available information.

Definition 1.2.4 (Natural filtration)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $T = \mathbb{Z}, \mathbb{N}$, $\mathbb{R}$ or $[0, \infty)$. Let $(X_t)_{t\in T}$ be a random process. The* natural filtration *of the process $(X_t)_{t\in T}$ is the the smallest filtration with respect to which $(X)_{t\in T}$ is adapted,*

$$\mathcal{F}_t^X = \sigma(X_s :\ s\ \leq t)\ ,\quad t \in T\ .$$

By definition, a stochastic process is adapted to its natural filtration.

1.2.2 Finite dimensional distributions

Given two measurable spaces $(\mathsf{X}_1, \mathcal{X}_1)$ et $(\mathsf{X}_2, \mathcal{X}_2)$, one defines the product measurable space $(\mathsf{X}_1 \times \mathsf{X}_2, \mathcal{X}_1 \otimes \mathcal{X}_2)$ where $\times$ denotes the Cartesian product of sets and $\otimes$ the corresponding product for σ-field: $\mathcal{X}_1 \otimes \mathcal{X}_2$ is the smallest σ-field containing the set class $\{A_1 \times A_2, A_1 \in \mathcal{X}_1 : A_2 \in \mathcal{X}_2\}$, which will be written

$$\mathcal{X}_1 \otimes \mathcal{X}_2 = \sigma\left(A_1 \times A_2 : A_1 \in \mathcal{X}_1, A_2 \in \mathcal{X}_2\right)\ .$$

Since the set class $\{A_1 \times A_2 : A_1 \in \mathcal{X}_1, A_2 \in \mathcal{X}_2\}$ is stable under finite intersections, a probability measure on $\mathcal{X}_1 \otimes \mathcal{X}_2$ is uniquely defined by its restriction to this class by Theorem B.1.5.

Similarly one defines a finite product measurable space $(\mathsf{X}_1 \times \cdots \times \mathsf{X}_n, \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n)$ from n measurable spaces $(\mathsf{X}_t, \mathcal{X}_t)$, $t \in T$. We will also write $(\prod_{t\in T} \mathsf{X}_t, \bigotimes_{t\in T} \mathcal{X})$.

Let now $(X_t)_{t\in T}$ be random process $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$ and $I \in \mathcal{I}$, where $\mathcal{I}$ denotes the set of finite subsets of T. Let $\mathbb{P}^{X_I}$ denotes the probability distribution of the random vector $\{X_t, t \in I\}$, that is, the image measure of $\mathbb{P}$ defined on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ by

$$\mathbb{P}^{X_I}\left(\prod_{t\in I} A_t\right) = \mathbb{P}\left(X_t \in A_t,\, t \in I\right)\ , \tag{1.1}$$

where A_t, $t \in T$ are any sets of the σ-field $\mathcal{X}$. The probability measure $\mathbb{P}^{X_I}$ is a *finite dimensional* distribution.

Definition 1.2.5. *We call* finite dimensional distributions *or* fidi distributions *of the process X the collection of probability measures $\left(\mathbb{P}^{X_I}\right)_{I\in\mathcal{I}}$.*

If T is infinite, the definition of product σ-fields above is extended by considering the σ-field generated by the *cylinders* on the Cartesian product $\prod_{t\in T} \mathsf{X}_t$ defined as the set of T-indexed sequences $(x_t)_{t\in T}$ such that $x_t \in \mathsf{X}_t$ for all $t \in T$. Let us focus on the case where

$(\mathsf{X}_t, \mathcal{X}_t) = (\mathsf{X}, \mathcal{X})$ for all $t \in T$. Then $\mathsf{X}^T = \prod_{t\in T} \mathsf{X}$ is the set of sequences $(x_t)_{t\in T}$ such that $x_t \in \mathsf{X}$ for all $t \in T$ and

$$\mathcal{X}^{\otimes T} = \sigma\left(\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I} : I \in \mathcal{I}, \forall t \in I,\ A_t \in \mathcal{X}\right) . \tag{1.2}$$

Then we have the following theorem.

Theorem 1.2.1. *Let $(X_t)_{t\in T}$ be random process $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$. Then the mapping $X : \omega \mapsto (X_t(\omega))_{t\in T}$ is measurable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$. Moreover the push forward measure $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$ is entirely characterized by the collection of finite distributions $\left(\mathbb{P}^{X_I}\right)_{I\in\mathcal{I}}$.*

Proof. Let us denote

$$\mathcal{C} = \left\{\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I} : I \in \mathcal{I}, \forall t \in I,\ A_t \in \mathcal{X}\right\} .$$

This class of sets is a π-system and $\mathcal{X}^{\otimes T} = \sigma(\mathcal{C})$. Moreover it is easy to show that, for all $A \in \mathcal{C}$, $X^{-1}(A) \in \mathcal{F}$ since it is a finite intersection of sets of the form $X_t^{-1}(A_t)$ which are in $\mathcal{F}$. And for the same reason, $\mathbb{P}(X^{-1}(A))$ is determined by $\mathbb{P}^{X_I}$ for a well chosen finite set I. Hence by Theorem B.1.5, $\mathbb{P}^X$ is uniquely determined by the collection of fidi distributions. □

This theorem thus shows that a process $(X_t)_{t\in T}$ can be seen as random variable $X = (X_t)_{t\in T}$ valued $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, whose *law* is determined by the fidi distributions.

Conversely, the fidi distributions can be obtain form $\mathbb{P}^X$ since the canonical projection Π_I of X^T on X^I defined by

$$\Pi_I(x) = (x_t)_{t\in I} \quad \text{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T , \tag{1.3}$$

is easily shown to be measurable from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ for all $I \subset T$. Hence, for all $I \subset T$,

$$\mathbb{P}^{(X_t)_{t\in I}} = \mathbb{P}^X \circ \Pi_I^{-1} .$$

Now, given a distribution $\mathbb{P}$ directly on the space of paths $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, it is also convenient to define a process on this measurable space, whose distribution will be given by $\mathbb{P}$. This construction is known as the *canonical process*.

Definition 1.2.6 (Canonical process)**.** *Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ the measurable space of corresponding paths. The canonical functions defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ is the collection of measurable functions $(\xi_t)_{t\in T}$ defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ and valued in $(\mathsf{X}, \mathcal{X})$ as $\xi_t(\omega) = \omega_t$ for all $\omega = (\omega_t)_{t\in t} \in \mathsf{X}^T$.*

When $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ is endowed with a probability measure $\mathbb{P}$, then the canonical process *$(\xi_t)_{t\in T}$ defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T}, \mathbb{P})$ has distribution $\mathbb{P}$.*

So far we have considered that the random process $X = (X_t)_{t\in T}$ is given as defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The only case where we have precised how all X_t are defined is in Definition 1.2.6, where we only need to have a probability $\mathbb{P}$ on the space $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to construct the whole process with the desired distribution. And we have seen that having this distribution is equivalent to have all the fidi distributions.

In practice, fidi distributions are much easier to deal with. Thus the following question arises:

Given a collection of distributions $(\nu_I)_{I\in\mathcal{I}}$ such that for all $I \in \mathcal{I}$, ν_I is a probability measure on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$, can we define a probability $\mathbb{P}$ on $(\mathsf{X}^T, \mathcal{X}^{\otimes T}$ such that $(\nu_I)_{I\in\mathcal{I}}$ are the fidi distributions associated to $\mathbb{P}$, $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$ for all $I \in \mathcal{I}$?

To answer this question. Suppose first that $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$ for all $I \in \mathcal{I}$. Let $J \subset I$ two finite subsets. Let us denote by $\Pi_{I,J}$ the canonical projection of X^I onto X^J defined by

$$\Pi_{I,J}[x] = (x_t)_{t\in J} \quad \text{for all} \quad x = (x_t)_{t\in I} \in \mathsf{X}^I \, . \tag{1.4}$$

The canonical projection is only preserving the vector entries that correspond to the indices of the subset J. Then, since $\Pi_J = \Pi_{I,J} \circ \Pi_I$, we get that $\mathbb{P} \circ \Pi_J^{-1} = \mathbb{P} \circ \Pi_I^{-1} \circ \Pi_{I,J}^{-1}$ and thus

$$\nu_J = \nu_I \circ \Pi_{I,J}^{-1} \qquad \text{for all } J \subset I \, . \tag{1.5}$$

This relationship is the formal translation of the fact that the fidi dimensional distribution of $J \subset I$ is obtained from that of I by integrating with respect to the variables X_t, where t belongs to the complementary set of J in I. This property induce a particular structure in the collection of fidi distributions. In particular, they must at least satisfy the compatibility condition (1.5). We shall soon see that this condition is in fact sufficient.

The following theorem shows how from the collection of all fidi distributions, one can get back to a unique probability measure on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, provided that Condition (1.5) holds and that $(\mathsf{X}, \mathcal{X})$ satisfies some topological conditions, such as $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ (for more general topological spaces, in particular infinite dimensional ones, see [Dudley, 2002, Theorem 12.1.2])). Note that only the existence part of this result is new as the uniqueness part is already a consequence of Theorem 1.2.1.

Theorem 1.2.2 (Kolmogorov)**.** *Set $\mathsf{X} = \mathbb{R}^p$ and $\mathcal{X} = \mathcal{B}(\mathbb{R}^p)$ for some $p \geq 1$. Let $\mathcal{I}$ be the set of all finite subsets of T. Let $(\nu_I)_{I\in\mathcal{I}}$ be such that, for all $I \in \mathcal{I}$, ν_I is a probability measure on $(\mathsf{X}^I, \mathcal{X}^{\otimes I})$ and satisfies Condition* (1.5)*. Then there exists a unique probability measure $\mathbb{P}$ on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ such that, for all $I \in \mathcal{I}$, $\nu_I = \mathbb{P} \circ \Pi_I^{-1}$.*

The following example is a simple application of this theorem.

Example 1.2.1 (Process of independent random variables)**.** *Let $(\nu_t)_{t\in T}$ be a collection of probability measures on $(\mathsf{X}, \mathcal{X})$. For all $I \in \mathcal{I}$, set*

$$\nu_I = \bigotimes_{t\in I} \nu_t \, , \tag{1.6}$$

where $\otimes$ denotes the tensor product of measures, that is, ν_I is the distribution of a vector with independent entries and marginal distributions given by ν_t, $t \in I$. It is easy to see that one defines a compatible collection of probability measures $(\nu_I)_{I\in\mathcal{I}}$ in the sense that Condition (1.5) holds. Hence, setting $\Omega = \mathsf{X}^T$, $X_t(\omega) = \omega_t$ and $\mathcal{F} = \sigma(X_t, t \in T)$, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $(X_t)_{t\in T}$ is a collection of independent random variables and $X_t \sim \nu_t$ for all $t \in T$. We say that $(X_t)_{t\in T}$ is a process of independent random variables (sometimes shortened as an independent process*) with marginal distributions $(\nu_t)_{t\in T}$. If moreover all ν_t's are equal to the same ν, we say that $(X_t)_{t\in T}$ is a process of independent and identically distributed (i.i.d.) random variables (sometimes shortened as an* i.i.d. process*) with marginal distribution ν.*

1.3 Gaussian processes

We now introduce an important class of random processes that can be seen as an extension of Gaussian vectors to the infinite-dimensional case. Let us recall first the definition of Gaussian random variables, univariate and then multivariate. More details can be found in [Jacod and Protter, 2003, Chapter 16].

1.3.1 Gaussian vectors

We start with $\mathbb{R}$-valued Gaussian variables.

Definition 1.3.1 (Gaussian variable)**.** *A real valued random variable X is said to be Gaussian if there exists $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ such that its characteristic function satisfies :*

$$\phi_X(u) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX}\right] = \mathrm{e}^{\mathrm{i}\mu u - \sigma^2 u^2/2} . \tag{1.7}$$

In this case, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Immediate properties are given in Exercise 1.1. In particular, one can show that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Thus, if $\sigma = 0$, then $X = \mu$ a.s. If $\sigma \neq 0$, then X admits the following probability density function

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{1.8}$$

Definition 1.3.1 can be extended to random vectors as follows.

Definition 1.3.2 (Gaussian vector)**.** *A random vector $[X_1, \ldots, X_n]^T$ valued in $\mathbb{R}^n$ is called a Gaussian vector if any linear combination of $X_1, \ldots, X_n$ is a Gaussian variable.*

Let μ denote the mean vector of $[X_1, \ldots, X_n]^T$ and Γ its covariance matrix. Then, for all $u \in \mathbb{R}^n$, the random variable $Y = \sum_{k=1}^n u_k X_k = u^T X$ is Gaussian. It follows that its distribution is determined by its mean and variance which can be expressed as

$$\mathbb{E}[Y] = \sum_{k=1}^n u_k \mathbb{E}[X_k] = u^T\mu \quad \text{and} \quad \mathrm{Var}(Y) = \sum_{j,k=1}^n u_j u_k \,\mathrm{Cov}(X_j, X_k) = u^T\Gamma u$$

Thus, the characteristic function of $[X_1, \ldots, X_n]^T$ can be written using μ and Γ as

$$\phi_X(u) = \mathbb{E}\left[\exp(\mathrm{i}u^T X)\right] = \mathbb{E}\left[\exp(\mathrm{i}Y)\right] = \exp\left(\mathrm{i}u^T\mu - \frac{1}{2}u^T\Gamma u\right) \tag{1.9}$$

Conversely, if a n-dimensional random vector X has a characteristic function of this form, we immediately obtain that X is a Gaussian vector from the characteristic function of its scalar products. This property yields the following proposition.

Proposition 1.3.1. *The probability distribution of an n-dimensional Gaussian vector X is determined by its mean vector and covariance matrix Γ. We will denote*

$$X \sim \mathcal{N}(\mu, \Gamma) .$$

Conversely, for all vector $\mu \in \mathbb{R}^n$ and all non-negative definite symmetric matrix $\Gamma \in \mathbb{R}^{n\times n}$, the distribution $X \sim \mathcal{N}(\mu, \Gamma)$ is well defined.

Proof. The first part of the result follows directly from (1.9). It also yields the following lemma.

Lemma 1.3.2. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ a non-negative definite symmetric matrix. Then for all $p \times n$ matrix A and $\mu' \in \mathbb{R}^n$, we have $\mu' + AX \sim \mathcal{N}(\mu' + A\mu, A\Gamma A^T)$.*

Let us now show the second (converse) part. First it holds for $n = 1$ as we showed previously. The case where Γ is diagonal follows easily. Indeed, let σ_i^2, $i = 1, \ldots, n$ denote the diagonal entries of Γ and set $\mu = [\mu_1, \ldots, \mu_n]^T$. Then take $X_1, \ldots, X_n$ independent such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. We then get $X \sim \mathcal{N}(\mu, \Gamma)$ by writing its characteristic function. To conclude the proof of Proposition 1.3.1, just observe that all non-negative definite symmetric matrix Γ can be written as $\Gamma = U\Sigma U^T$ with Σ diagonal with non-negative entries and U orthogonal. Thus taking $Y \sim \mathcal{N}(0, \Sigma)$ and setting $X = \mu + UY$, the above lemma implies that $X \sim \mathcal{N}(\mu, \Gamma)$, which concludes the proof. □

The following proposition is easy to get (see [Jacod and Protter, 2003, Corollaire 16.1]).

Proposition 1.3.3. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ a non-negative definite symmetric matrix. Then X has independent components if and only if Γ is diagonal.*

Using the same path as in the proof of Proposition 1.3.1, i.e. by considering the cases where Γ is diagonal and using the diagionalization in an orthogonal basis to get the general case, one gets the following result (see [Jacod and Protter, 2003, Corollaire 16.2]).

Proposition 1.3.4. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ a non-negative definite symmetric matrix. If Γ is full rank, the probability distribution of X admits a density defined in $\mathbb{R}^n$ by*

$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Gamma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right), \quad x \in \mathbb{R}^n .$$

If Γ's rank $r < n$, that is, Γ has an $n-r$-dimensional null space, X belongs, with probability 1, to an r-dimensional affine subspace of $\mathbb{R}^n$. Indeed, there are $r - n$ linearly independent vectors a_i such that $\mathrm{Cov}(a_i^T X) = 0$ and thus $a_i^T X = a_i^T \mu$ a.s. Obviously X does not admit a density function in this case.

1.3.2 Real valued Gaussian processes

Having recalled the classical results on Gaussian vectors, we now introduce the definition of *Gaussian processes*.

Definition 1.3.3 (Gaussian processes)**.** *A real-valued random process $X = (X_t)_{t\in T}$ is called a Gaussian process if, for all finite set of indices $I = \{t_1, t_2, \cdots, t_n\}$, $[X_{t_1}, X_{t_2}, \cdots, X_{t_n}]^T$ is a Gaussian vector.*

Thus a Gaussian vector $[X_1, \ldots, X_n]^T$ may itself be seen as a Gaussian process $(X_t)_{t\in\{1,\ldots,n\}}$. This definition therefore has an interest in the case where T has an infinite cardinality. According to (1.9), the collection of fidi distributions is characterized by the mean function $\mu : t \in T \mapsto \mu(t) \in \mathbb{R}$ and the covariance function $\gamma : (t, s) \in (T \times T) \mapsto \gamma(t, s) \in \mathbb{R}$. Moreover, for all finite set of indices $I = \{t_1, t_2, \ldots, t_n\}$, the matrix Γ_I with entries $\Gamma_I(m, k) = \gamma(t_m, t_k)$,

with $1 \leq m, k \leq n$, is a covariance matrix of a random vector of dimension n. It is therefore nonnegative definite symmetric. Conversely, given a function $\mu : t \in T \mapsto m(t) \in \mathbb{R}$ and a function $\gamma : (t,s) \in (T \times T) \mapsto \gamma(t,s) \in \mathbb{R}$ such that, for all finite set of indices I, the matrix Γ_I is nonnegative definite symmetric, we can define, for all finite set of indices $I = \{t_1, t_2, \cdots, t_n\}$, a Gaussian probability ν_I on $\mathbb{R}^n$ by

$$\nu_I \stackrel{\text{def}}{=} \mathcal{N}(\mu_I, \Gamma_I) \tag{1.10}$$

where $\mu_I = [\mu(t_1), \ldots, \mu(t_n)]^T$. The so defined collection $(\nu_I)_{I \in \mathcal{I}}$, satisfies the compatibility conditions and this implies, by Theorem 1.2.2, the following result.

Theorem 1.3.5. *Let T be any set of indices, μ a real valued function defined on T and γ a real valued function defined on $T \times T$ such that all restrictions Γ_I to the set $I \times I$ with $I \subseteq T$ finite are nonnegative definite symmetric matrices. Then one can define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a Gaussian process $(X_t)_{t \in T}$ defined on this space with* mean μ *and* covariance function γ*, that is such that, for all s, $t \in T$,*

$$\mu(t) = \mathbb{E}[X_t] \quad \textit{and} \quad \gamma(s,t) = \mathbb{E}[(X_s - \mu(s))(X_t - \mu(t))] \ .$$

As a consequence we can extend the usual notation $\mathcal{N}(\mu, \gamma)$ as follows.

Definition 1.3.4 (Gaussian process fidi distributions)**.** *Let T be any index set. Let μ be any real valued function on T and γ any real valued function defined on $T \times T$ satisfying the condition of Theorem 1.3.5. We denote by $\mathcal{N}(\mu, \gamma)$ the law of the Gaussian process with mean μ and covariance γ in the sense of fidi distributions.*

1.4 Stopping Times

In this section, we consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbb{N}}, \mathbb{P})$ and an adapted process $((X_n, \mathcal{F}_n))_{n \in \mathbb{N}}$. The σ-field $\mathcal{F}_\infty$ is defined as the smallest one containing all $\mathcal{F}_k$, $k \in \mathbb{N}$, that is

$$\mathcal{F}_\infty = \bigvee_{k \in \mathbb{N}} \mathcal{F}_k \ .$$

In many examples, $(\mathcal{F}_n)_{n \in \mathbb{N}}$ is the natural filtration of some given process $(Y_m)_{m \in \mathbb{N}}$.

The term *stopping time* is an expression from gambling. A game of chance which evolves in time (for example an infinite sequence of coin tosses) can be adequately represented by a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbb{N}}, \mathbb{P})$, the sub-$\sigma$-fields $\mathcal{F}_n$ giving the information on the results of the game available to the player at time n. A stopping rule for the player thus consists of giving a rule for leaving the game at time n, based at each time on the information at his disposal at that time. The time τ of stopping the game by such a rule is a stopping time. Note that stopping times may take the value $+\infty$, corresponding to the case where the game does not stop.

Definition 1.4.1 (Stopping times)**.** *A random variable τ from Ω to $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ is called a stopping time on the filtered measurable space $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k \in \mathbb{N}})$ if, for all $k \in \mathbb{N}$, $\{\tau \leq k\} \in \mathcal{F}_k$. The family $\mathcal{F}_\tau$ of events $A \in \mathcal{F}_\infty$ such that, for every $k \in \mathbb{N}$, $A \cap \{\tau \leq k\} \in \mathcal{F}_k$, is called the σ-field of events prior to time τ.*

It can be easily checked that $\mathcal{F}_\tau$ is indeed a σ-field (Exercise 1.5). Since $\{\tau = n\} = \{\tau \leq n\} \setminus \{\tau \leq n-1\}$, one can replace $\{\tau \leq n\}$ by $\{\tau = n\}$ in the definition of the stopping time τ and in the definition of the σ-field $\mathcal{F}_\tau$. It may sometimes be useful to note that the constant random variables are also stopping times. In such a case, there exists some $n \in \mathbb{N}$ such that $\tau(\omega) = n$ for every $\omega \in \Omega$, and $\mathcal{F}_\tau = \mathcal{F}_n$.

For any stopping time τ, the event $\{\tau = \infty\}$ belongs to $\mathcal{F}_\infty$, for it is the complement of the union of the event $\{\tau = n\}$, $n \in \mathbb{N}$, which all belong to $\mathcal{F}_\infty$. It follows that $B \cap \{\tau = \infty\} \in \mathcal{F}_\infty$ for all $B \in \mathcal{F}_\tau$, showing that $\tau : \Omega \to \bar{\mathbb{N}}$ is $\mathcal{F}_\infty$ measurable.

Definition 1.4.2 (Hitting times)**.** *For $A \in \mathcal{X}$, the first hitting time τ_A and σ_A of the set A are the $\bar{\mathbb{N}}$-valued random-variables respectively defined on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ by*

$$\tau_A : (x_n)_{n\in\mathbb{N}} \mapsto \inf\{n \geq 0 \,:\, x_n \in A\} \ ,$$
$$\sigma_A : (x_n)_{n\in\mathbb{N}} \mapsto \inf\{n \geq 1 \,:\, x_n \in A\} \ ,$$

where, by convention, $\inf \emptyset = +\infty$. The successive positive hitting times $\sigma_A^{(n)}$, $n \geq 0$, are defined inductively by $\sigma_A^{(0)} = 0$ and for all $k \geq 0$,

$$\sigma_A^{(k+1)} : x = (x_n)_{n\in\mathbb{N}} \mapsto \inf\left\{n > \sigma_A^{(k)}(x) \,:\, x_n \in A\right\} \ . \tag{1.11}$$

Hitting times are stopping times with respect to the canonical filtration, since, for all $n \in \mathbb{N}$,

$$\{\tau_A \leq n\} = \bigcup_{k=0}^{n} \{\xi_k \in A\} \in \mathcal{F}_n \ ,$$

and

$$\{\sigma_A \leq 0\} = \emptyset \quad \text{and if } n \geq 1 \,,\ \{\sigma_A \leq n\} = \bigcup_{k=1}^{n} \{\xi_k \in A\} \in \mathcal{F}_n^\xi \ .$$

The stopping times property is preserved through many standard operations, as shown in the following useful result.

Proposition 1.4.1. *Let $(\Omega,\ \mathcal{F},\ (\mathcal{F}_k)_{k\in\mathbb{N}},\ \mathbb{P})$ be a filtered probability space and τ and σ be two stopping times for the filtration $(\mathcal{F}_n)_{n\geq 0}$. Denote by $\mathcal{F}_\tau$ and $\mathcal{F}_\sigma$ the σ-fields of the events prior to τ and σ, respectively. Then,*

(i) $\tau \wedge \sigma$, $\tau \vee \sigma$ and $\tau + \sigma$ are stopping times,

(ii) if $\tau \leq \sigma$, then $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$,

(iii) $\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$,

(iv) $\{\tau < \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$, $\{\tau = \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$.

Proof.

(i) Let $n \in \mathbb{N}$. We show that the events $\{\tau \wedge \sigma \leq n\}$, $\{\tau \vee \sigma \leq n\}$ and $\{\tau + \sigma \leq n\}$ belong to $\mathcal{F}_n$. Since

$$\{\tau \wedge \sigma \leq n\} = \{\tau \leq n\} \cup \{\sigma \leq n\}$$

and τ and σ are stopping times, $\{\tau \leq n\}$ and $\{\sigma \leq n\}$ belong to $\mathcal{F}_n$; therefore $\{\tau \wedge \sigma \leq n\} \in \mathcal{F}_n$. Similarly, $\{\tau \vee \sigma \leq n\} = \{\tau \leq n\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$. Finally,

$$\{\tau + \sigma \leq n\} = \bigcup_{k=0}^{n} \{\tau \leq k\} \cap \{\sigma \leq n-k\} \; .$$

Now, for $0 \leq k \leq n$, $\{\tau \leq k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\sigma \leq n-k\} \in \mathcal{F}_{n-k} \subset \mathcal{F}_n$; hence $\{\tau + \sigma \leq n\} \in \mathcal{F}_n$.

(ii) Let $A \in \mathcal{F}_\tau$ and $n \in \mathbb{N}$. As $\{\sigma \leq n\} \subset \{\tau \leq n\}$, $A \cap \{\sigma \leq n\} = A \cap \{\tau \leq n\} \cap \{\sigma \leq n\}$. Now $A \cap \{\tau \leq n\} \in \mathcal{F}_n$ and $\{\sigma \leq n\} \in \mathcal{F}_n$ (σ is a stopping time); therefore $A \cap \{\tau \leq n\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$ and $A \cap \{\sigma \leq n\} \in \mathcal{F}_n$. Thus $A \in \mathcal{F}_\sigma$.

(iii) It follows from (ii) that $\mathcal{F}_{\tau\wedge\sigma} \subset \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. Conversely, let $A \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. Obviously $A \subset \mathcal{F}_\infty$. To prove that $A \in \mathcal{F}_{\tau\wedge\sigma}$, one must show that, for every $k \geq 0$, $A \cap \{\tau \wedge \sigma \leq k\} \in \mathcal{F}_k$. We have $A \cap \{\tau \leq k\} \in \mathcal{F}_k$ and $A \cap \{\sigma \leq k\} \in \mathcal{F}_k$. Hence, since $\{\tau \wedge \sigma \leq k\} = \{\tau \leq k\} \cup \{\sigma \leq k\}$, we get

$$A \cap \{\tau \wedge \sigma \leq k\} = A \cap (\{\tau \leq k\} \cup \{\sigma \leq k\}) = (A \cap \{\tau \leq k\}) \cup (A \cap \{\sigma \leq k\}) \in \mathcal{F}_k \; .$$

(iv) Let $n \in \mathbb{N}$. We write

$$\{\tau < \sigma\} \cap \{\tau \leq n\} = \bigcup_{k=0}^{n} \{\tau = k\} \cap \{\sigma > k\} \; .$$

Now, for $0 \leq k \leq n$, $\{\tau = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\sigma > k\} = \{\sigma \leq k\}^c \in \mathcal{F}_k \subset \mathcal{F}_n$. Therefore, $\{\tau < \sigma\} \cap \{\tau \leq n\} \in \mathcal{F}_n$, showing that $\{\tau < \sigma\} \in \mathcal{F}_\tau$. Similarly,

$$\{\tau < \sigma\} \cap \{\sigma \leq n\} = \bigcup_{k=0}^{n} \{\sigma = k\} \cap \{\tau < k\}$$

and since, for $0 \leq k \leq n$, $\{\sigma = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{\tau < k\} = \{\tau \leq k-1\} \in \mathcal{F}_{k-1} \subset \mathcal{F}_n$, it also holds $\{\tau < \sigma\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$ so that $\{\tau < \sigma\} \in \mathcal{F}_\sigma$. Finally, $\{\tau < \sigma\} \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma$. The last statement of the proposition follows from

$$\{\tau = \sigma\} = \{\tau < \sigma\}^c \cap \{\sigma < \tau\}^c \in \mathcal{F}_\tau \cap \mathcal{F}_\sigma \; .$$

□

1.5 Exercises

Exercise 1.1. Let $X_0 \sim \mathcal{N}(0,1)$ as in Definition 1.3.1 and let $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.

1. Show that $\mu + \sigma X_0 \sim \mathcal{N}(\mu, \sigma^2)$.
2. Show that the characteristic function ϕ_{X_0} of X_0 is infinitely differentiable and that all its derivatives are integrable (w.r.t. the Lebesgue measure on $\mathbb{R}$).
3. Deduce that X_0 admits a density and that $\mathbb{E}[X_0^p]$ is finite for all $p \in \mathbb{N}$.

4. Deduce $\mathbb{E}[X_0]$ and $\mathbb{E}\left[X_0^2\right]$.

Let us define

$$g(u) \stackrel{\text{def}}{=} \int \mathrm{e}^{\frac{u^2}{2}-\frac{x^2}{2}+\mathrm{i}u\,x}\,\mathrm{d}x$$

5. Show that g is well defined and continuously differentiable on $\mathbb{R}$.

6. Show that g is constant on $\mathbb{R}$ and compute $g(0)$ by writing $g(0)^2$ as an integral on $\mathbb{R}^2$, that can be computed using polar coordinates.

7. Deduce that X_0 has density $p_{0,1}$ as defined by (1.8).

8. Conclude that a variable with distribution $\mathcal{N}(\mu,\sigma^2)$ has mean μ and variance σ^2, and that, if $\sigma^2>0$, it admits the density given in (1.8).

Exercise 1.2. Let X be a Gaussian vector, A_1 and A_2 two linear applications. Let us set $X_1 = A_1X$ and $X_2 = A_2X$. Give the distribution of (X_1, X_2) and a necessary and sufficient condition for X_1 and X_2 to be independent.

Exercise 1.3. Let X be a Gaussian random variable, with zero mean and unit variance, $X \sim \mathcal{N}(0,1)$. Let $Y = X\mathbf{1}_{\{U=1\}} - X\mathbf{1}_{\{U=0\}}$ where U is a Bernoulli random variable with parameter $1/2$ independent of X. Show that $Y \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}(X,Y)=0$ but also that X and Y are not independent.

Exercise 1.4. Let $n \geq 1$ and Γ be a $n \times n$ nonnegative definite hermitian matrix.

1. Find a Gaussian vector X valued in $\mathbb{R}^n$ and a unitary matrix U such that UX has covariance matrix Γ. [Hint : take a look at the proof of Proposition 1.3.1].

2. Show that
$$\Sigma := \frac{1}{2}\begin{bmatrix}\mathrm{Re}(\Gamma) & -\mathrm{Im}(\Gamma)\\ \mathrm{Im}(\Gamma) & \mathrm{Re}(\Gamma)\end{bmatrix}$$
is a real valued $(2n)\times(2n)$ nonnegative definite symmetric matrix.

Let X and Y be two n-dimensional Gaussian vectors such that

$$\begin{bmatrix}X & Y\end{bmatrix}^T \sim \mathcal{N}(0,\Sigma)\ .$$

3. What is the covariance matrix of $Z = X + \mathrm{i}Y$?

4. Compute $\mathbb{E}\left[ZZ^T\right]$.

The random variable Z is called a centered circularly-symmetric normal vector.

Let now T be an arbitrary index set, $\mu : I \to \mathbb{C}$ and $\gamma : T^2 \to \mathbb{C}$ such that for all finite subset $I \subset T$, the matrix $\Gamma_I = [\gamma(s,t)]_{s,t\in I}$ is a nonnegative definite hermitian matrix.

5. Use the previous questions to show that there exists a random process $(X_t)_{t\in T}$ valued in $\mathbb{C}$ such that, for all $s,t \in T$,
$$\mathbb{E}[X_t] = \mu(t) \quad \text{and} \quad \mathrm{Cov}(X_s,X_t) = \gamma(s,t)\ .$$

Exercise 1.5. Let τ be a stopping time on the filtered measurable space $(\Omega,\mathcal{F},(\mathcal{F}_k)_{k\in\mathbb{N}})$. Show that $\mathcal{F}_\tau$ in Definition 1.4.1 is a sub-σ-field of $\mathcal{F}$.

Exercise 1.6. Let τ be a $(\mathcal{F}_n)$-stopping time on $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Let Y be a real random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Show that Y is $\mathcal{F}_\tau$-measurable if and only if $Y\mathbb{1}_{\{\tau \leq n\}}$ is $\mathcal{F}_n$-measurable for all $n \in \bar{\mathbb{N}}$.
2. Show that $A \in \mathcal{F}_\tau$ if and only if $A \cap \{\tau = n\} \in \mathcal{F}_n$ for all $n \in \bar{\mathbb{N}}$.
3. Let $n \in \mathbb{N}$ and Y be a real random variable $\mathcal{F}_n$-measurable. Prove that $Y\mathbb{1}_{\{\tau=n\}}$ is $\mathcal{F}_\tau$-measurable.
4. Let X be an integrable real random variable. Show that for all $n \in \overline{\mathbb{N}}$,

$$\mathbb{E}\left[X\mathbb{1}_{\{\tau=n\}} \middle| \mathcal{F}_\tau\right] = \mathbb{1}_{\{\tau=n\}}\mathbb{E}\left[X \middle| \mathcal{F}_n\right] .$$

Chapter 2

Weakly stationary time series

2.1 Strict stationarity of a random process in discrete time

2.1.1 Definition

Stationarity plays a central role in stochastic modelling. We will distinguish two versions of this property, *strict stationarity* which says that the distribution of the random process is invariant by shifting the time origin and a *weak stationarity*, which imposes that only the first and second moments are invariant, with the additional assumption that these moments exist.

Definition 2.1.1 (Shift and backshift operators)**.** *Suppose that $T = \mathbb{Z}$ or $T = \mathbb{N}$. We denote by* S *and call the* shift operator *the mapping $\mathsf{X}^T \to \mathsf{X}^T$ defined by*

$$\mathrm{S}(x) = (x_{t+1})_{t\in T} \quad \textit{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T \ .$$

For all $\tau \in T$, we define S^τ by

$$\mathrm{S}^\tau(x) = (x_{t+\tau})_{t\in T} \quad \textit{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T \ .$$

The operator $\mathrm{B} = \mathrm{S}^{-1}$ is called the backshift operator*.*

Definition 2.1.2 (Strict stationarity)**.** *Set $T = \mathbb{Z}$ or $T = \mathbb{N}$. A random process $(X_t)_{t\in T}$ is strictly stationary if X and $\mathrm{S}\circ X$ have the same law, i.e. $\mathbb{P}^{\mathrm{S}\circ X} = \mathbb{P}^X$.*

Since the law is characterized by fidi distributions, one has $\mathbb{P}^{\mathrm{S}\circ X} = \mathbb{P}^X$ if and only if

$$\mathbb{P}^{\mathrm{S}\circ X} \circ \Pi_I^{-1} = \mathbb{P}^X \circ \Pi_I^{-1}$$

for all finite subset $I \in \mathcal{I}$. Now $\mathbb{P}^{\mathrm{S}\circ X} \circ \Pi_I^{-1} = \mathbb{P}^X \circ (\Pi_I \circ \mathrm{S})^{-1}$ and $\Pi_I \circ \mathrm{S} = \Pi_{I+1}$, where $I + 1 = \{t + 1, t \in I\}$. We conclude that $\{X_t, t \in T\}$ is *strictly stationary* if and only if, for all finite set $I \in \mathcal{I}$,

$$\mathbb{P}^{X_I} = \mathbb{P}_{I+1} \ .$$

Also observe that the strict stationarity implies that X and $\mathrm{S}^\tau \circ X$ has the same law for all $\tau \in T$ and thus $\mathbb{P}^{X_I} = \mathbb{P}_{I+\tau}$, where $I + \tau = \{t + \tau, t \in I\}$.

Example 2.1.1 (I.i.d process)**.** *Let $(Z_t)_{t\in T}$ be a sequence of* independent and identically distributed *(i.i.d) with values in $\mathbb{R}^d$. Then $(Z_t)_{t\in T}$ is a strictly stationary process , since, for all finite set $I = \{t_1, < t_2 < \cdots < t_n\}$ and all Borel set $A_1, \ldots, A_n$ of $\mathbb{R}^d$, we have*

$$\mathbb{P}(Z_{t_1} \in A_1, \cdots, Z_{t_n} \in A_n) = \prod_{j=1}^{n} \mathbb{P}(Z_0 \in A_j) \ ,$$

which does not depend on $t_1, \ldots, t_n$. Recall that, from Example 1.2.1, for all probability ν on $\mathbb{R}^d$, we can define a random process $(Z_t)_{t\in T}$ which is i.i.d. with marginal distribution *ν, that is, such that $Z_t \sim \nu$ for all $t \in T$.*

2.1.2 Stationarity preserving transformations

In this section, we set $T = \mathbb{Z}$, $\mathsf{X} = \mathbb{C}^d$ et $\mathcal{X} = \mathcal{B}(\mathbb{C}^d)$ for some integer $d \geq 1$. Let us start with an illustrating example.

Example 2.1.2 (Moving transformation of an i.i.d. process)**.** *Let Z be an i.i.d. process (see Example 2.1.1). Let k be an integer and g a measurable function from $\mathbb{R}^k$ to $\mathbb{R}$. One can check that the process $(X_t)_{t\in\mathbb{Z}}$ defined by*

$$X_t = g(Z_t, Z_{t-1}, \cdots, Z_{t-k+1})$$

also is a stationary random process in the strict sense. On the other hand, the obtained process is not i.i.d. in general since for $k \geq 1$, $X_t, X_{t+1}, \ldots, X_{t+k-1}$ are identically distributed but are in general dependent variables as they all depend on the same random variables Z_t. Nevertheless such a process is said to be k-dependent *because $(X_s)_{s\leq t}$ and $(X_s)_{s>t+k}$ are independent for all t. The m-dependent processes can be used to approximate a large class of dependent processes to study the asymptotic behavior of statistics such as usual the sample mean.*

Observe that in this example, to derive the stationarity of X, it is not necessary to use that Z is i.i.d., only that it is stationary. In fact, to check stationarity, it is often convenient to reason directly on the laws of the trajectories using the notion of filtering.

Definition 2.1.3. *Let ϕ be a measurable function from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{Y}^T, \mathcal{Y}^{\otimes T})$ and $X = (X_t)_{t\in T}$ be a process with values in $(\mathsf{X}, \mathcal{X})$. A ϕ- filtering with input X and output Y means that the random process $Y = (Y_t)_{t\in T}$ is defined as $Y = \phi \circ X$, or, equivalently, $Y_t = \Pi_t(\phi(X))$ for all $t \in T$, where Π_t is a shorthand notation for $\Pi_{\{t\}}$ defined in (1.3). Thus Y takes its values in $(\mathsf{Y}, \mathcal{Y})$. If ϕ is linear, we will say that Y is obtained by* linear filtering *of X.*

In Example 2.1.2, X is obtained by ϕ-filtering Z (non-linearly, unless g is a linear form) with $\phi : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ defined by

$$\phi\big((x_t)_{t\in\mathbb{Z}}\big) = \big(g(x_t, x_{t-1}, \ldots, x_{t-k+1})\big)_{t\in\mathbb{Z}} \ .$$

Example 2.1.3 (Shift)**.** *A very basic linear filtering is obtained with $\phi = \mathrm{S}$ where S is the shift operator of Definition 2.1.1. In this case $Y_t = X_{t+1}$ for all $t \in \mathbb{Z}$.*

Example 2.1.4 (Finite impulse response filter (FIR)). *Let $n \geq 1$ and $t_1 < \cdots < t_n$ in $\mathbb{Z}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$. Then $\phi = \sum_i \alpha_i \mathrm{S}^{-t_i}$ defines a linear filtering and for any input $X = (X_t)_{t\in\mathbb{Z}}$, the output is given by*

$$Y_t = \sum_{i=1}^{n} \alpha_i X_{t-t_i}, \quad t \in \mathbb{Z}\,.$$

Example 2.1.5 (Differencing operator). *A particular case is the* differencing operator $\mathbf{1}-\mathrm{S}^{-1}$ *where* $\mathbf{1}$ *denotes the identity on* X^T. *The output then reads as*

$$Y_t = X_t - X_{t-1}, \quad t \in \mathbb{Z}\,.$$

One can iterate this operator so that $Y = (\mathbf{1} - \mathrm{S}^{-1})^k X$ is given by

$$Y_t = \sum_{j=0}^{k} \binom{k}{j} (-1)^j X_{t-j}, \quad t \in \mathbb{Z}\,.$$

Example 2.1.6 (Time reversion). *Let $X = \{X_t, t \in \mathbb{Z}\}$ be a random process. Time reversion then set the output as*

$$Y_t = X_{-t}, \quad t \in \mathbb{Z}\,.$$

Note that in all previous examples the operators introduced preserve the strict stationarity, that is to say, if the input X is strictly stationary then so is the output Y. It is easy to construct a linear filtering which does not preserve the strict stationarity, for example, $y = \phi(x)$ with $y_t = x_t$ for t even and $Y_t = x_t + 1$ for t odd. A property stronger than the conservation of stationarity and very easy to verify is given by the following definition.

Definition 2.1.4. *A ϕ-filter is* shift invariant *if ϕ commutes with* S, $\phi \circ \mathrm{S} = \mathrm{S} \circ \phi$.[1]

It is easy to show that a shift-invariant filter preserves the strict stationarity. However it is a stronger property. The time reversion is an example of a filter that is not shift-invariant, although it does preserve the strict stationarity. Indeed, in this case, we have $\phi \circ \mathrm{S} = \mathrm{S}^{-1} \circ \phi$. All the other examples above are shift-invariant.

Remark 2.1.1. *A* shift invariant *ϕ-filter is entirely determined by its composition with the canonical projection Π_0 defined as in (1.3) but with $I = \{0\}$. Indeed, let $\phi_0 = \Pi_0 \circ \phi$. Then for all $s \in \mathbb{Z}$, $\Pi_s \circ \phi = \Pi_0 \circ \mathrm{S}^s \circ \phi = \Pi_0 \circ \phi \circ \mathrm{S}^s$. Since for all $x \in \mathsf{X}^T$, $\phi(x)$ is the sequence $(\Pi_s \circ \phi)_{s\in T}$, we get the result.*

In this chapter, we focus on second order properties of univariate time series, that is, on their means and covariance functions. It turns out that the stationarity induces a particular structure of the covariances of a time series that can be exploited to provide a spectral representation of the time series. Finally we will conclude the chapter with the Wold decomposition, which basically shows that any weakly stationary processes, up to an additive deterministic-like component, can be expressed by linearly filtering a white noise (the innovation process).

[1] There is a slight hidden discrepancy in this definition: if ϕ is defined from $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ to $(\mathsf{Y}^T, \mathcal{Y}^{\otimes T})$ with $\mathsf{X} \neq \mathsf{Y}$ then the notation S refers to two different shifts: one on X^T and the other one on Y^T.

2.2 L^2 processes

The definition of the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$, or simply L^2, is recalled in Example A.1.4.

Definition 2.2.1 (L^2 univariate time series)**.** *The process $X = (X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{C}$ is an L^2 process if $X_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ for all $t \in T$.*

The *mean function* defined on T by $\mu(t) = \mathbb{E}[X_t]$ takes its values in $\mathbb{C}$ and the *covariance function* is defined on $T \times T$ by

$$\gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathbb{E}\left[(X_s - \mu(s))\overline{(X_t - \mu(t))}\right] ,$$

which takes its values in $\mathbb{C}$. We will sometimes use the notation μ_X and γ_X, the subscript X indicating the process used in these definitions. For all $s \in T$, $\gamma(s,s)$ is a variance and is thus nonnegative. More generally, the following properties hold.

Proposition 2.2.1. *Let Γ be the covariance function of a L^2 process $X = (X_t)_{t\in T}$ with values in $\mathbb{C}^d$. The following properties hold.*

(i) Hermitian symmetry: for all $s, t \in T$,

$$\gamma(s,t) = \overline{\gamma(t,s)} \tag{2.1}$$

(ii) Nonnegativity: for all $n \geq 1$, $t_1, \ldots, t_n \in T$ and $a_1, \cdots, a_n \in \mathbb{C}$,

$$\sum_{1\leq k,m\leq n} \overline{a_k}\gamma(t_k, t_m)a_m \geq 0 \tag{2.2}$$

Conversely, if γ satisfy these two properties, there exists an L^2 process $X = (X_t)_{t\in T}$ with values in $\mathbb{C}$ with covariance function γ.

Proof. Relation (2.1) is immediate. To show (2.2), define the linear combination $Y = \sum_{k=1}^n \overline{a_k} X_{t_k}$. Y is a complex valued random variable. Using that the Cov operator is hermitian, we get

$$\mathrm{Var}(Y) = \sum_{1\leq k,m\leq n} \overline{a_k}\gamma(t_k, t_m)a_m$$

which implies (2.2).

The converse assertion follows from Exercise 1.4. □

2.3 Univariate weakly stationary time series

From now on, in this chapter, we take $T = \mathbb{Z}$. If an L^2 process is strictly stationary, then its first and second order properties must satisfy certain properties. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a strictly stationary L^2 process with values in $\mathbb{C}$. Then its mean function is constant, since its marginal distribution is invariant. Moreover its covariance function Γ satisfies $\gamma(s,t) = \gamma(s-t, 0)$ for all $s, t \in \mathbb{Z}$ since the bi-dimensional marginals are also invariant by a translation of time. A weakly stationary process inherits these properties but is not necessary strictly stationary, as in the following definition.

Definition 2.3.1 ((Univariate) weakly stationary time series)**.** *Let $\mu \in \mathbb{C}$ and $\gamma : \mathbb{Z} \to \mathbb{C}$. A process $(X_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}$ is said* weakly stationary *with mean μ and autocovariance function γ if all the following assertions hold:*

(i) X is an L^2 process, i.e. $\mathbb{E}\left[|X_t|^2\right] < +\infty$,

(ii) for all $t \in \mathbb{Z}$, $\mathbb{E}\left[X_t\right] = \mu$,

(iii) for all $(s,t) \in \mathbb{Z} \times \mathbb{Z}$, $\mathrm{Cov}\left(X_s, X_t\right) = \gamma(s-t)$.

By definition the autocovariance function of a weakly stationary process is defined on T instead of T^2 for the covariance function in the general case.

As already mentioned a strictly stationary L^2 process is weakly stationary. The converse implication is of course not true in general. It is true however for Gaussian processes defined in Section 1.3, see Proposition 1.3.1.

Here we considered time series valued in $\mathbb{C}$, hence called univariate. We will also consider weakly stationary time series valued in a general Hilbert space. In the case where this space is $\mathbb{C}^d$, the obtained time series is called *multivariate*.

2.3.1 Properties of the autocovariance function

The properties of Proposition 2.2.1 imply the following ones in the case of a weakly stationary process.

Proposition 2.3.1. *The autocovariance function $\gamma : \mathbb{Z} \to \mathbb{C}$ of a complex valued weakly stationary process satisfies the following properties.*

(i) Hermitian symmetry : for all $s \in \mathbb{Z}$,

$$\gamma(-s) = \overline{\gamma(s)}$$

(ii) Nonnegative definiteness : for all integer $n \geq 1$ and $a_1, \cdots, a_n \in \mathbb{C}$,

$$\sum_{s=1}^{n}\sum_{t=1}^{n} \overline{a_s}\gamma(s-t)a_t \geq 0$$

The autocovariance matrix Γ_n of n consecutive samples $X_1, \ldots, X_n$ of the time series has a particular structure, namely it is constant on its diagonals, $(\Gamma_n)_{ij} = \gamma(i-j)$,

$$\begin{aligned}\Gamma_n &= \mathrm{Cov}([X_1 \ \ldots \ X_n]^T) \\ &= \begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(1-n) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(2-n) \\ \vdots & & & \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix}\end{aligned} \tag{2.3}$$

On says that Γ_n is a *Toeplitz* matrix . Since $\gamma(0)$ is generally non-zero (note that otherwise X_t is zero a.s. for all t), it can be convenient to normalize the autocovariance function in the following way.

Definition 2.3.2 (Autocorrelation function)**.** *Let X be a weakly stationary process with autocovariance function γ such that $\gamma(0) \neq 0$. The* autocorrelation function *of X is defined as*

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \quad \tau \in \mathbb{Z} .$$

It is normalized in the sense that $\rho(0) = 1$ and $|\rho(s)| \leq 1$ for all $s \in \mathbb{Z}$.

The last assertion follows from the Cauchy-Schwarz inequality (see Theorem A.1.1),

$$|\gamma(s)| = |\mathrm{Cov}\,(X_s, X_0)| \leq \sqrt{\mathrm{Var}\,(X_s)\,\mathrm{Var}\,(X_0)} = \gamma(0) ,$$

the last equality following from the weakly stationary assumption.

Let us give some simple examples of weakly stationary processes. We first examine a very particular case.

Definition 2.3.3 (White noise)**.** *A weak white noise is a centered weakly stationary process whose autocovariance function satisfies $\gamma(0) = \sigma^2 > 0$ and $\gamma(s) = 0$ for all $s \neq 0$. We will denote $(X_t) \sim \mathrm{WN}(0, \sigma^2)$. When a weak white noise is an i.i.d. process, it is called a* strong white noise. *We will denote $(X_t) \sim \mathrm{IID}(0, \sigma^2)$.*

Of course a strong white noise is a weak white noise. However the converse is in general not true. The two definitions only coincide for Gaussian processes because in this case the independence is equivalent to being uncorrelated.

Example 2.3.1 (MA(1) process)**.** *Define, for all $t \in \mathbb{Z}$,*

$$X_t = Z_t + \theta Z_{t-1} , \tag{2.4}$$

where $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ and $\theta \in \mathbb{R}$. Then $\mathbb{E}[X_t] = 0$ and the autocovariance function reads

$$\gamma(s) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } s = 0, \\ \sigma^2\theta & \text{if } s = \pm 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Such a weakly stationary process is called a Moving Average of order 1 MA(1).

Example 2.3.2 (Harmonic process)**.** *Let $(A_k)_{1\leq k\leq N}$ be N real valued L^2 random variables. Denote $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$. Let $(\Phi_k)_{1\leq k\leq N}$ be N i.i.d. random variables with a uniform distribution on $[-\pi, \pi]$, and independent of $(A_k)_{1\leq k\leq N}$. Define*

$$X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k) , \tag{2.6}$$

where $(\lambda_k)_{1\leq k\leq N} \in [-\pi, \pi]$ are N frequencies. The process (X_t) is called an harmonic process. *It satisfies $\mathbb{E}[X_t] = 0$ and, for all $s, t \in \mathbb{Z}$,*

$$\mathbb{E}[X_s X_t] = \frac{1}{2} \sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k (s-t)) .$$

It is thus a weakly stationary process.

Example 2.3.3 (Random walk). *Let $(S_t)_{t\in\mathbb{N}}$ be a random process defined by $S_0 = 0$ and, for all $t \in \mathbb{N}^*$, by $S_t = X_1 + \cdots + X_t$, where (X_t) is a strong white noise with variance σ^2. Such a process is called a* random walk. *We have $\mathbb{E}[S_t] = 0$, $\mathbb{E}\left[S_t^2\right] = t\sigma^2$ and for all $s \leq t \in \mathbb{N}$,*

$$\mathbb{E}[S_s S_t] = \mathbb{E}[(S_s + X_{s+1} + \cdots + X_t)S_s] = s\,\sigma^2$$

The process $(S_t)_{t\in\mathbb{N}}$ is not weakly stationary.

Example 2.3.4 (Continued from Example 2.3.1). *Consider the function χ defined on $\mathbb{Z}$ by*

$$\chi(s) = \begin{cases} 1 & \text{if } s = 0, \\ \rho & \text{if } s = \pm 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

where $\rho \in \mathbb{R}$. It is the autocovariance function of a real valued process if and only if $\rho \in [-1/2, 1/2]$. We know from Example 2.3.1 that χ is the autocovariance function of a real valued MA(1) process if and only if $\sigma^2(1+\theta^2) = 1$ and $\sigma^2\theta = \rho$ for some $\theta \in \mathbb{R}$. If $|\rho| \leq 1/2$, the solutions to this equation are

$$\theta = (2\rho)^{-1}(1 \pm \sqrt{1 - 4\rho^2}) \quad \text{and} \quad \sigma^2 = (1+\theta^2)^{-1}\ .$$

If $|\rho| > 1/2$, there are no real solutions. In fact, in this case, it can even be shown that there is no real valued weakly stationary process whose autocovariance is χ, see Exercise 2.4.

Some simple transformations of processes preserve the weak stationarity. Linearity is crucial in this case since otherwise the second order properties of the output cannot solely depend on the second order properties of the input.

Example 2.3.5 (Invariance of the autocovariance function under time reversion (continued from Example 2.1.6). *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a weakly stationary process with mean μ_X and autocovariance function γ_X. Denote, for all $t \in \mathbb{Z}$, $Y_t = X_{-t}$ as in Example 2.1.6. Then (Y_t) is weakly stationary with same mean as X and autocovariance function $\gamma_Y = \overline{\gamma_X}$.*

$$\mathbb{E}[Y_t] = \mathbb{E}[X_{-t}] = \mu_X\ ,$$
$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mathrm{Cov}(X_{-t-h}, X_{-t}) = \gamma_X(-h) = \overline{\gamma_X(h)}\ .$$

2.3.2 Empirical mean and autocovariance function

Suppose that we observe n consecutive samples of a real valued weakly stationary time series $X = (X_t)$. Can we have a rough idea of the second order parameters of X μ and γ ? This is an estimation problem. The first step for answering this question is to provide estimators of μ and γ. Since these quantities are defined using an expectation $\mathbb{E}$, a quite natural approach is to replace this expectation by an empirical sum over the observed data. This yields the *empirical mean*

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} X_k\ , \tag{2.8}$$

and the *empirical autocovariance* and *autocorrelation* functions

$$\hat{\gamma}_n(h) = \frac{1}{n}\sum_{k=1}^{n-|h|}(X_k - \hat{\mu}_n)(X_{k+|h|} - \hat{\mu}_n) \quad \text{and} \quad \hat{\rho}_n(h) = \hat{\gamma}_n(h)/\hat{\gamma}_n(0)\ . \tag{2.9}$$

Let us examine how such estimators look like on some examples.

Example 2.3.6 (Heartbeats (Continued from Example 1.1.1))**.** *Take the data displayed in Figure 1.1, which roughly looks stationary. Its empirical autocorrelation is displayed in Figure 2.1. We observe a positive correlation in the sense that the obtained values are significantly above the x-axis, at least if one compares with the empirical correlation obtained from a sample of a Gaussian white noise with the same length.*

A positive autocorrelation $\rho(h)$ has a simple interpretation: it means that X_t and X_{t+h} have a tendency of being on the same side of their means with a higher probability. A more precise interpretation is to observe that, in the sense of the Hilbert space L^2 (see Appendix A),

$$\mathrm{proj}\,(X_{t+h} - \mu | \,\mathrm{Span}\,(X_t - \mu)) = \rho(h)(X_t - \mu)\,,$$

and the error has variance $\gamma(0)(1 - |\rho(h)|^2)$ (see Exercise 2.6). In practice, we do not have access to the exact computation of these quantities from a single sample $X_1, \ldots, X_n$. We can however let t varies at fixed h, hoping that the evolution in t more or less mimic the variation in ω. In Figure 2.2, we plot X_t VS X_{t+1} and indeed see this phenomenon: $\hat{\rho}(1) = 0.966$ indicate that X_{t+1} is very well approximated by a linear function of X_t, as can be observed in this figure.

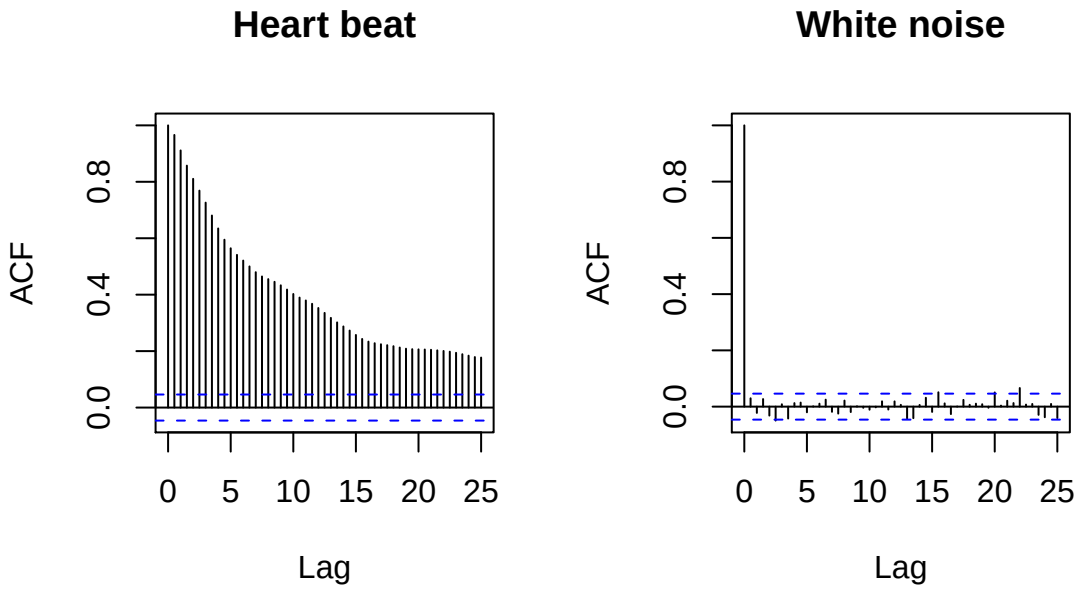

Figure 2.1: *Left : empirical autocorrelation $\hat{\rho}_n(h)$ of heartbeat data for $h = 0, \ldots, 100$. Right : the same from a simulated white noise sample with same length.*

2.4 Spectral measure

Recall that $\mathbb{T}$ denotes any interval congruent to $[0, 2\pi)$. We denote by $\mathcal{B}(\mathbb{T})$ the associated Borel σ-field. The Herglotz theorem shows that the autocovariance function of a weakly

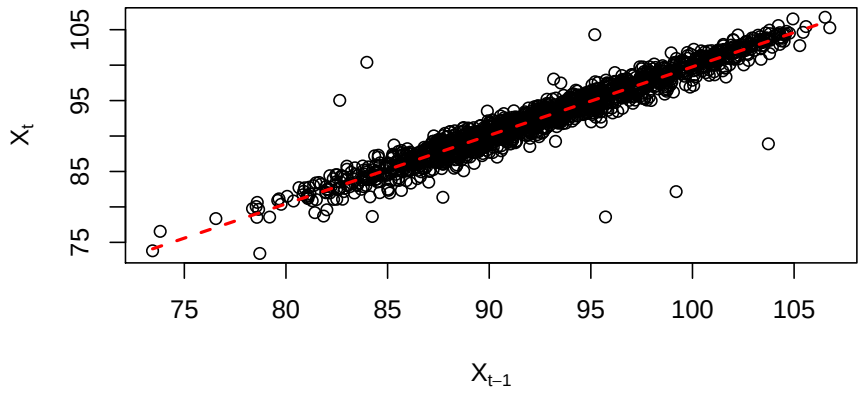

Figure 2.2: *Each point is a couple* (X_{t-1}, X_t)*, where* $X_1, \ldots, X_n$ *is the heartbeat data sample. The dashed line is the best approximation of* X_t *as a linear function of* X_{t-1}*.*

stationary process X is entirely determined by a finite nonnegative measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. This measure is called the *spectral measure* of X.

Theorem 2.4.1 (Herglotz)**.** *A sequence* $(\gamma(h))_{h\in\mathbb{Z}}$ *is a nonnegative definite hermitian sequence in the sense of Proposition 2.3.1 if and only if there exists a finite nonnegative measure* ν *on* $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ *such that :*

$$\gamma(h) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu(\mathrm{d}\lambda), \ \forall h \in \mathbb{Z}\,. \tag{2.10}$$

Moreover this relation defines ν *uniquely.*

Remark 2.4.1. *By Proposition 2.3.1, Theorem 2.4.1 applies to all* γ *which is an autocovariance function of a weakly stationary process* X*. In this case* ν *(also denoted* ν_X*) is called the* spectral measure *of* X*. If* ν *admits a density* f*, it is called the* spectral density function*.*

Proof. Suppose first that $\gamma(n)$ satisfies (2.10) with ν as in the theorem. Then γ is an hermitian function. Let us show it is a nonnegative definite hermitian function. Fix a positive integer n. For all $a_k \in \mathbb{C}$, $1 \le k \le n$, we have

$$\sum_{k,m} a_k \overline{a_m} \gamma(k-m) = \int_{\mathbb{T}} \sum_{k,m} a_k \overline{a_m} \mathrm{e}^{\mathrm{i}k\lambda} \mathrm{e}^{-im\lambda} \nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} \left| \sum_k a_k \mathrm{e}^{\mathrm{i}k\lambda} \right|^2 \nu(\mathrm{d}\lambda) \ge 0\,.$$

Hence γ is nonnegative definite.

Conversely, suppose that γ is a nonnegative definite hermitian sequence. For all $n \ge 1$,

define the function

$$f_n(\lambda) = \frac{1}{2\pi n}\sum_{k=1}^{n}\sum_{m=1}^{n}\gamma(k-m)\mathrm{e}^{-\mathrm{i}k\lambda}\mathrm{e}^{\mathrm{i}m\lambda}$$
$$= \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma(k)\mathrm{e}^{-\mathrm{i}k\lambda}\ .$$

Since γ is nonnegative definite, we get from the first equality that $f_n(\lambda) \geq 0$, for all $\lambda \in \mathbb{T}$. Define ν_n as the nonnegative measure with density f_n on $\mathbb{T}$. We get that

$$\int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\nu_n(\mathrm{d}\lambda) = \int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}f_n(\lambda)\mathrm{d}\lambda = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma(k)\int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}(h-k)\lambda}\mathrm{d}\lambda$$
$$= \begin{cases}\left(1-\frac{|h|}{n}\right)\gamma(h), & \text{if } |h|<n\ ,\\ 0, & \text{otherwise}\ .\end{cases} \tag{2.11}$$

We can multiply the sequence (ν_n) by a constant to obtain a sequence of probability measures. Thus Theorem C.2.3 implies that there exists a nonnegative measure ν and a subsequence (ν_{n_k}) of (ν_n) such that

$$\lim_{k\to\infty}\int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\nu_{n_k}(\mathrm{d}\lambda) = \int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\nu(\mathrm{d}\lambda),\ .$$

Using (2.11) and taking the limit of the subsequence, we get that

$$\gamma(h) = \int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\nu(\mathrm{d}\lambda),\ \forall h\in\mathbb{Z}\ .$$

Let us conclude with the uniqueness of ν. Suppose that another nonnegative measure ξ satisfies for all $h \in \mathbb{Z}$: $\int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\nu(\mathrm{d}\lambda) = \int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}h\lambda}\mu(\mathrm{d}\lambda)$. Then by Theorem A.3.1, we obtain that $\int_{\mathbb{T}} g(\lambda)\nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} g(\lambda)\mu(\mathrm{d}\lambda)$ for all continuous (2π)-periodic function g. This implies $\nu = \mu$. □

Corollary 2.4.2 (The ℓ^2 case)**.** *Let $(\gamma(h))_{h\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. Then it is a nonnegative definite hermitian sequence in the sense of Proposition 2.3.1 if and only if*

$$f(\lambda) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}}\gamma(h)\mathrm{e}^{-\mathrm{i}h\lambda}\ ,$$

where the convergence holds in $L^2(\mathbb{T})$, is nonnegative for almost every λ.

Proof. It suffices to apply Theorem 2.4.1 and Corollary A.3.3. □

The proof also shows that f is the spectral density function associated to γ.

Example 2.4.1 (MA(1), Continued from Example 2.3.4)**.** *Consider Example 2.3.4. Then $(\chi(h))$ is in $\ell^1(\mathbb{Z})$ and*

$$f(\lambda) = \frac{1}{2\pi}\sum_{h}\chi(h)\mathrm{e}^{-\mathrm{i}h\lambda} = \frac{1}{2\pi}(1+2\rho\cos(\lambda))\ .$$

Thus we obtain that χ is nonnegative definite if and only if $|\rho| \leq 1/2$. An exemple of such a spectral density function is displayed in Figure 2.3.

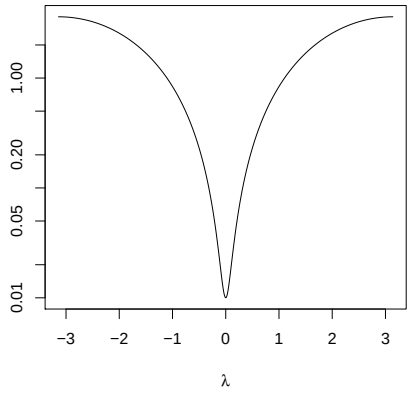

Figure 2.3: *Spectral density function (in logarithmic scale) of an MA(1) process, as given by (2.4) with* $\sigma = 1$ *and* $\theta = -0.9$.

Example 2.4.2 (Spectral density function of a white noise)**.** *Recall the definition of a white noise, Definition 2.3.3. We easily get that the white noise* $\mathrm{IID}(0, \sigma^2)$ *admits a spectral density function given by*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \, ,$$

that is, a constant spectral density function. Hence the name "white noise", referring to white color that corresponds to a constant frequency spectrum.

Example 2.4.3 (Spectral measure of an harmonic process, continued from Example 2.3.2)**.** *The autocovariance function of* X *is given by (see Example 2.3.2)*

$$\gamma(h) = \frac{1}{2} \sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k h) \, , \tag{2.12}$$

where $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$*. Observing that*

$$\cos(\lambda_k h) = \frac{1}{2} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}h\lambda} (\delta_{\lambda_k}(\mathrm{d}\lambda) + \delta_{-\lambda_k}(\mathrm{d}\lambda))$$

where $\delta_{x_0}(\mathrm{d}\lambda)$ *denote the Dirac mass at point* x_0*, the spectral measure of* X *reads*

$$\nu(\mathrm{d}\lambda) = \frac{1}{4} \sum_{k=1}^{N} \sigma_k^2 \delta_{\lambda_k}(\mathrm{d}\lambda) + \frac{1}{4} \sum_{k=1}^{N} \sigma_k^2 \delta_{-\lambda_k}(\mathrm{d}\lambda) \, .$$

We get a sum of Dirac masses with weights σ_k^2 *and located at the frequencies of the harmonic functions.*

Harmonic processes have singular properties. The autocovariance function in (2.12) implies that covariance matrices Γ_n are expressed as a sum of $2N$ matrices with rank 1. Thus Γ_n is not invertible as soon as $n > 2N$ and thus harmonic process fall in the following class of process.

Definition 2.4.1 (Finitely linearly predictable processes)**.** *A centered weakly stationary process X is called* finitely linearly predictable *if there exists $n \geq 1$ such that for all $t \geq n$, $X_t \in \mathrm{Span}\,(X_1, \ldots, X_n)$ (in the L^2 sense).*

One can wonder whether the other given examples are linearly predictable. The answer is given by the following result, whose proof is left to the reader (see Exercise 2.9).

Proposition 2.4.3. *Let γ be the autocovariance function of a centered weakly stationary process X. If $\gamma(0) \neq 0$ and $\gamma(t) \to 0$ as $t \to \infty$ then X is not finitely linearly predictable.*

2.5 Spectral representations of weakly stationary processes

2.5.1 Orthogonally scattered random measures

In this section, we let $(\mathbb{X}, \mathcal{X})$ denote any measurable space.

Definition 2.5.1 (Orthogonally scattered random measures)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An* orthogonally scattered random measure *W on $(\mathbb{X}, \mathcal{X})$ is a L^2 random process indexed on $\mathcal{X}$, say $W = (W(A))_{A \in \mathcal{X}}$ such that*

(i) For all $A \in \mathcal{X}$, $\mathbb{E}\,[W(A)] = 0$.

(ii) For all $A, B \in \mathcal{X}$ such that $A \cap B = \emptyset$, $W(A)$ and $W(B)$ are uncorrelated and $W(A \cup B) = W(A) + W(B)$;

(iii) For all nonincreasing sequence $(A_n)_{n \in \mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$, we have $\mathrm{Var}\,(W(A_n)) \to 0$.

Lemma 2.5.1. *Let W an orthogonally scattered random measure on $(\mathbb{X}, \mathcal{X})$. Let $A \in \mathcal{X}$, and set $\eta(A) = \mathrm{Var}\,(W(A))$. Then η is a finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$. Moreover, for all $A, B \in \mathcal{X}$, $\mathrm{Cov}\,(W(A), W(B)) = \eta(A \cap B)$.*

Proof. To show that η is a measure, it is sufficient to show that η is additive and continuous, that is, for all nonincreasing sequence $(A_n)_{n \in \mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$, we have $\eta(A_n) = 0$. These two properties follow from (ii) and (iii) of Definition 2.5.1.

Observe that $A = (A \setminus B) \cup (A \cap B)$ and $B = (B \setminus A) \cup (A \cap B)$ and that $A \setminus B$, $B \setminus A$ and $A \cap B$ are disjoint sets. By (ii) in Definition 2.5.1, we have $W(A) = W(A \setminus B) + W(A \cap B)$ and $W(B) = W(B \setminus A) + W(A \cap B)$ and, moreover, $W(A \setminus B)$, $W(B \setminus A)$ are $W(A \cap B)$ are uncorrelated. Consequently,

$$\mathrm{Cov}\,(W(A), W(B)) = \mathrm{Var}\,(W(A \cap B)) = \eta(A \cap B)\ ,$$

which concludes the proof. □

The measure η is called the *intensity measure* of W. The previous lemma comes with the converse following result.

Lemma 2.5.2. *Let W be a L^2 random process indexed by $\mathcal{X}$ such that, for all $A \in \mathcal{X}$, $\mathbb{E}\,[W(A)] = 0$. Suppose that there exists a measure η on $(\mathbb{X}, \mathcal{X})$ such that, for all $A, B \in \mathcal{X}$, $\mathrm{Cov}\,(W(A), W(B)) = \eta(A \cap B)$. Then W is an orthogonally scattered random measure on $(\mathbb{X}, \mathcal{X})$ with intensity measure η.*

Proof. Let $A, B \in \mathcal{X}$ such that $A \cap B = \emptyset$. We have that

$$\begin{aligned}\mathrm{Var}\left(W(A \cup B) - W(A) - W(B)\right) \\ = \eta(A \cup B) + \eta(A) + \eta(B) - 2\eta(A) - 2\eta(B) + 2\eta(A \cap B) = 0 \ ,\end{aligned}$$

where we used $\eta(A \cup B) = \eta(A) + \eta(B)$ and $\eta(A \cap B) = 0$. Thus $W(A \cup B) = W(A) + W(B)$ and the additivity property (ii) is satisfied.

Consider a nonincreasing sequence $(A_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ such that $\bigcap_{n=0}^{\infty} A_n = \emptyset$. Since η is a measure, we have $\mathrm{Var}\left(W(A_n)\right) = \eta\left(A_n\right) \to 0$ which gives (iii). Hence the result. □

Example 2.5.1 (Randomly weighted sum of Dirac masses)**.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a centered complex valued L^2 random process defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Denote for all $n \geq 0$, $\sigma_n^2 = \mathrm{Var}\left(Y_n\right)$ and assume that $(\sigma_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ and $\mathrm{Cov}\left(Y_n, Y_k\right) = 0$ for $n \neq k$. Let $(\lambda_n)_{n\in\mathbb{N}} \subset \mathbb{X}$. Define the process W indexed by $\mathcal{X}$ as*

$$W = \sum_{n=0}^{\infty} Y_n \delta_{\lambda_n} \ ,$$

where δ_x is the Dirac mass at x. Then, for all $A, B \in \mathcal{X}$,

$$\mathrm{Cov}\left(W(A), W(B)\right) = \sum_{n=0}^{\infty} \sigma_n^2 \mathbb{1}_A(\lambda_n)\mathbb{1}_B(\lambda_n) = \eta(A \cap B) \ ,$$

where

$$\eta(A) = \sum_{n=0}^{\infty} \sigma_n^2 \delta_{\lambda_n}(A) \ .$$

By Lemma 2.5.2, W is an orthogonally scattered random measure on $(\mathbb{X}, \mathcal{X})$ with intensity measure η.

2.5.2 Stochastic integral

Theorem 2.5.3. *Let W be an orthogonally scattered random measure on $(\mathbb{X}, \mathcal{X})$ with intensity measure η. Then there exists a unique isometric operator* w *from $L^2(\mathbb{X}, \mathcal{X}, \eta)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathrm{w}(\mathbb{1}_A) = W(A)$ for all $A \in \mathcal{X}$.*

For all $f \in L^2(\mathbb{X}, \mathcal{X}, \eta)$, we further have $\mathbb{E}\left[\mathrm{w}(f)\right] = 0$ and we have

$$\mathrm{w}(L^2(\mathbb{X}, \mathcal{X}, \eta)) = \overline{\mathrm{Span}}\left(W(A), A \in \mathcal{X}\right) \ ,$$

where the closure is understood in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Proof. Set $\mathcal{H} = L^2(\mathbb{X}, \mathcal{X}, \eta)$ and $\mathcal{I} = L^2(\Omega, \mathcal{F}, \mathbb{P})$. For $A, B \in \mathcal{X}$, we have

$$\langle \mathbb{1}_A, \mathbb{1}_B \rangle_{\mathcal{H}} = \int \mathbb{1}_A \mathbb{1}_B \mathrm{d}\eta = \langle W(A), W(B) \rangle_{\mathcal{I}} \ .$$

Since by Proposition A.2.4, we have

$$\overline{\mathrm{Span}}\left(\mathbb{1}_A, A \in \mathcal{X}\right) = L^2(\mathbb{X}, \mathcal{X}, \eta) \ ,$$

the result follows from the extension theorem for isometric operators (see Theorem A.6.2). □

It can be convenient to use the same notation for W and the isometric operator w. Since $f \mapsto \mathrm{w}(f)$ is a linear operator on a set of unctions, it is also common to use the integral notation although this integral relies on a particular L^2 construction (and thus do not satisfy the usual nice properties of the classical integral),

$$\int f \, \mathrm{d}W = \int f(\lambda) \, \mathrm{d}W(\lambda) = W(f) = \mathrm{w}(f) \; , \tag{2.13}$$

which satisfies, for all $(f, g) \in L^2(\mathbb{X}, \mathcal{X}, \eta)$ and $(u, v) \in \mathbb{C} \times \mathbb{C}$,

$$\int (uf + vg) \, \mathrm{d}W = u \int f \, \mathrm{d}W + v \int g \, \mathrm{d}W \; .$$

and

$$\mathbb{E}\left[\left(\int f \, \mathrm{d}W \right) \overline{\left(\int g \, \mathrm{d}W \right)} \right] = \int f \bar{g} \, \mathrm{d}\eta \; .$$

Moreover; since $\mathbb{E}\left[W(A)\right] = 0$ for all $A \in \mathcal{X}$ and $\mathbb{E}$ is continuous on $L^2(\mathbb{X}, \mathcal{X}, \eta)$, we have

$$\mathbb{E}\left[\int f \, \mathrm{d}W \right] = 0 \; .$$

We will call $\int f \, \mathrm{d}W$ the *stochastic integral* of f with respect to W.

Interestingly, for any finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$, all isometric operators from $L^2(\mathbb{X}, \mathcal{X}, \nu)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ can be interpreted as a stochastic integral with intensity measure ν.

Theorem 2.5.4. *Let ν be a finite nonnegative measure on $(\mathbb{X}, \mathcal{X})$ and* J *an isometric operator from $L^2(\mathbb{X}, \mathcal{X}, \nu)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that for all $f \in L^2(\mathbb{X}, \mathcal{X}, \nu)$, $\mathbb{E}\left[\mathrm{J}(f)\right] = 0$. Then there exists an orthogonally scattered random measure W on $\mathbb{X}$ with intensity measure ν such that, for all $f \in L^2(\mathbb{X}, \mathcal{X}, \nu)$, $\mathrm{J}(f) = \int_{\mathbb{X}} f \, \mathrm{d}W$.*

Proof. To obtain W, we must set, for all $A \in \mathcal{X}$, $W(A) = \mathrm{J}(\mathbb{1}_A)$. Since J is an isometric operator, for all $A, B \in \mathcal{X}$,

$$\mathrm{Cov}\left(W(A), W(B)\right) = \left\langle \mathrm{J}(\mathbb{1}_A), \mathrm{J}(\mathbb{1}_B) \right\rangle_{L^2(\Omega, \mathcal{F}, \mathbb{P})} = \left\langle \mathbb{1}_A, \mathbb{1}_B \right\rangle_{L^2(\mathbb{X}, \mathcal{X}, \nu)} = \nu(A \cap B) \; .$$

and by Lemma 2.5.2, W is an orthogonally scattered random measure on $\mathbb{X}$ with intensity measure ν and it is the unique one whose integral coincides with J on $\{\mathbb{1}_A, \, A \in \mathcal{X}\}$. Since both are isometric operators and

$$\overline{\mathrm{Span}}\left(\mathbb{1}_A, \, A \in \mathcal{X}\right) = L^2(\mathbb{X}, \mathcal{X}, \nu) \; ,$$

they coincide on the whole space $L^2(\mathbb{X}, \mathcal{X}, \nu)$, which achieves the proof. □

2.5.3 Spectral representation and spectral domain

We now introduce the spectral representation associated to a weakly stationary process. It is based on an orthogonally scattered random measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. We first start from such a random measure.

Proposition 2.5.5. *Let W be an orthogonally scattered random measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ with intensity measure η. Then, the sequence $(X_t)_{t\in\mathbb{Z}}$ defined by*

$$X_t = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}t\lambda} \, \mathrm{d}W(\lambda) \ ,$$

is a centered weakly stationary process with spectral measure η.

Proof. Define $f_t(\lambda) = \mathrm{e}^{\mathrm{i}t\lambda}$ for all $t \in \mathbb{Z}$, so that $f_t \in L^2(\mathbb{T}, \eta)$. Since the stochastic integral is an isometric operator, we have, for all $(s,t) \in \mathbb{Z}^2$,

$$\mathrm{Cov}\,(X_s, X_t) = \mathbb{E}\left[X_s \bar{X}_t\right] = \langle W(f_s), W(f_t)\rangle_{L^2(\Omega,\mathcal{F},\mathbb{P})} = \langle f_s, f_t\rangle_{L^2(\mathbb{T},\eta)} = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}(s-t)\lambda}\eta(\mathrm{d}\lambda) \ .$$

Hence the result. □

Conversely, let us show that any centered weakly stationary process can be expressed as in Proposition 2.5.5.

Definition 2.5.2 (Linear closure of a random process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a L^2 process. Its* time domain*, denoted by $\mathcal{H}^X$ is defined as*

$$\mathcal{H}^X = \overline{\mathrm{Span}}\,(X_t,\, t \in \mathbb{Z}) \ ,$$

where the closure is understood in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

In other words, $\mathcal{H}^X$ is the space of all L^2 random variables that can be obtained as an L^2 limit of a sequence of finite linear combinations of $(X_t)_{t\in\mathbb{Z}}$.

Theorem 2.5.6 (Spectral representation)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with spectral measure ν. Then there exists an orthogonally scattered random measure $\hat{X}$ on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ with intensity measure ν such that, for all $t \in \mathbb{Z}$, we have the* spectral representation

$$X_t = \int \mathrm{e}^{\mathrm{i}t\lambda} \, \mathrm{d}\hat{X}(\lambda) \ .$$

Moreover, we call $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ the spectral domain *of X, and the mapping $f \mapsto \int f \, \mathrm{d}\hat{X}$ defines the unique unitary operator from the spectral domain $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ to the time domain $\mathcal{H}^X$ that maps each function $\lambda \mapsto \mathrm{e}^{\mathrm{i}t\lambda}$ to X_t.*

Proof. Set $\mathcal{H} = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$, $\mathcal{I} = L^2(\Omega, \mathcal{F}, \mathbb{P})$ and, for all $t \in \mathbb{Z}$, $f_t(\lambda) = \mathrm{e}^{\mathrm{i}t\lambda}$. We shall consider the sequences $(f_t)_{t\in\mathbb{Z}}$ and $(X_t)_{t\in\mathbb{Z}}$ in $\mathcal{H}$ and $\mathcal{I}$, respectively. By Corollary A.3.2, we have $\overline{\mathrm{Span}}\,(f_t,\ t \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ and by Theorem 2.4.1, for all $(s,t) \in \mathbb{Z}^2$,

$$\langle f_s, f_t\rangle_{L^2(\mathbb{T},\mathcal{B}(\mathbb{T}),\nu)} = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}s\lambda}\mathrm{e}^{-\mathrm{i}t\lambda}\nu(\mathrm{d}\lambda) = \mathrm{Cov}\,(X_s, X_t) = \langle X_s, X_t\rangle_{L^2(\Omega,\mathcal{F},\mathbb{P})} \ .$$

By Theorem A.6.2, there exists a unique isometric operator S_X from $\overline{\mathrm{Span}}\,(f_t,\ t \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu) = \mathcal{H}$ to $\mathcal{I}$ such that, for all $t \in \mathbb{Z}$, $\mathrm{S}_X(f_t) = X_t$. As an operator from $\mathcal{H}$ to $\mathrm{S}_X(\mathcal{H}) = \overline{\mathrm{Span}}\,(X_t,\, t \in \mathbb{Z}) = \mathcal{H}^X$, it is unitary (since then it is isometric and surjective).

Applying Theorem 2.5.4, we obtain the spectral representation based on $\hat{X}$, which satisfies all the claimed properties. This concludes the proof. □

The reader may legitimately wonder where all this abstract framework take us to. In fact it will be very useful to express standard linear filters on weakly stationary processes. The reason is the following. The spectral representation defined in Theorem 2.5.6 using $\hat{X}$ can be used to represent any random variable Y in the time domain $\mathcal{H}^X$ as the stochastic integral of a function $f \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$,

$$Y = \int f \, \mathrm{d}\hat{X} \, .$$

Moreover this defines f uniquely since we have a isometric isomorphism between the spectral domain and the time domain. It follows that a linear operator on $\mathcal{H}^X$ can equivalently be seen as a linear operator on $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$.

Let $H^X = \mathrm{Span}\,(X_t,\, t \in \mathbb{Z})$, so that its closure in L^2 is the time domain $\mathcal{H}^X$. Suppose that X is not linearly predictable, so that any element $Y \in H^X$ has a unique representation

$$Y = \sum_{t \in \mathbb{Z}} \lambda_t X_t \, ,$$

where $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ with finite support.

Take now a shift-invariant linear filter F on $\mathbb{C}^{\mathbb{Z}}$. A linear operator $\tilde{\mathrm{F}}$ is obtained on H^X by setting, for any $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ with finite support,

$$\tilde{\mathrm{F}}\left(\sum_{t\in\mathbb{Z}} \lambda_t X_t\right) = \Pi_0 \circ \mathrm{F}\left(\sum_{t\in\mathbb{Z}} \lambda_t S^t(X)\right) = \sum_{t\in\mathbb{Z}} \lambda_t \Pi_t \circ \mathrm{F}(X_t) \, ,$$

since $\Pi_0 \circ \mathrm{F} \circ S^t = \Pi_t \circ \mathrm{F}$. In other words, $\tilde{\mathrm{F}}(X_t) = \Pi_t \circ \mathrm{F}(X)$ and it is extended linearly to H^X. If $\tilde{\mathrm{F}}$ is continuous on H^X as a subset of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, then it admits a unique continuous extension to $\mathcal{H}^X$. Using the spectral representation, the operator $\tilde{\mathrm{F}}$ can as well be studied on $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$, in particular its continuity properties.

Unfortunately the confusion between the original operator F and its induction $\tilde{\mathrm{F}}$ on $\mathcal{H}^X$ is widespread in the literature on time series.

Example 2.5.2 (Backshift operator)**.** *Consider the backshift operator* $\mathrm{B} = \mathrm{S}^{-1}$*, see Definition 2.1.1. In this case, the induced operator on* H^X *is defined by*

$$\tilde{\mathrm{B}}(X_t) = X_{t-1} \, ,$$

At first sight, it seems hard to express this operator for any $Y \in \mathcal{H}^X$*. However using the spectral representation,*

$$Y = \int f \, \mathrm{d}\hat{X} \, ,$$

we immediately get

$$\tilde{\mathrm{B}}(Y) = \int f(\lambda) \mathrm{e}^{-\mathrm{i}\lambda} \, \hat{X}(\mathrm{d}\lambda) \, ,$$

first for any $f : \lambda \mapsto \mathrm{e}^{\mathrm{i}t\lambda}$ *with* $t \in \mathbb{Z}$*, then for* f *being a trigonometric polynomial* $\lambda \mapsto \sum_{t\in\mathbb{Z}} \lambda_t \mathrm{e}^{\mathrm{i}t\lambda}$*, where* $(\lambda_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$ *with finite support, and finally for any* $f \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$ *by continuity extension.*

2.6 Innovation process

In this section, we let $X = (X_t)_{t\in\mathbb{Z}}$ denote a centered weakly stationary processes. We shall define the Wold decomposition of X. This decomposition mainly relies on the concept of innovations. Let

$$\mathcal{H}_t^X = \overline{\mathrm{Span}}\,(X_s, s \le t)$$

denote the *linear past* of a given random process $X = (X_t)_{t\in\mathbb{Z}}$ up to time t. It is related to the already mentioned space $\mathcal{H}^X$ as follows

$$\mathcal{H}^X = \overline{\bigcup_{t\in\mathbb{Z}} \mathcal{H}_t^X}\ .$$

Let us introduce the *innovations* of a weakly stationary process.

Definition 2.6.1 (Innovation process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. We call* innovation process *the process $\epsilon = (\epsilon_t)_{t\in\mathbb{Z}}$ defined by*

$$\epsilon_t = X_t - \mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1}^X\right)\ . \tag{2.14}$$

By the orthogonal principle (see Theorem A.4.1), each ϵ_t is characterized by the fact that $X_t - \epsilon_t \in \mathcal{H}_{t-1}^X$ (which implies $\epsilon_t \in \mathcal{H}_t^X$) and $\epsilon_t \perp \mathcal{H}_{t-1}^X$. As a consequence $(\epsilon_t)_{t\in\mathbb{Z}}$ is a centered orthogonal sequence. We shall see below that it is in fact a white noise, that is, the variance of the innovation

$$\sigma^2 = \|\epsilon_t\|^2 = \mathbb{E}\left[|\epsilon_t|^2\right] \tag{2.15}$$

does not depend on t.

Example 2.6.1 (Innovation process of a white noise)**.** *The innovation process of a white noise $X \sim \mathrm{WN}(0, \sigma^2)$ is $\epsilon = X$.*

Example 2.6.2 (Innovation process of a MA(1), continued from Example 2.3.1)**.** *Consider the process X defined in Example 2.3.1. Observe that $Z_t \perp \mathcal{H}_{t-1}^X$. Thus, if $\theta Z_{t-1} \in \mathcal{H}_{t-1}^X$, we immediately get that $\epsilon_t = Z_t$. The questions are thus: is Z_{t-1} in $\mathcal{H}_{t-1}^X$? and, if not, what can be done to compute ϵ_t ?*

Because the projection in (2.14) is done on an infinite dimension space, it is interesting to compute it as a limit of finite dimensional projections. To this end, define, for $p \ge 0$, the finite dimensional space

$$\mathcal{H}_{t,p}^X = \mathrm{Span}\,(X_s, t - p < s \le t)\ ,$$

and observe that $(\mathcal{H}_{t,p}^X)_p$ is an increasing sequence of linear space whose union has closure $\mathcal{H}_t^X$. Hence by Property (ii) in Theorem A.4.3, we have, for any L^2 variable Y,

$$\lim_{p\to\infty} \mathrm{proj}\left(Y \middle| \mathcal{H}_{t,p}^X\right) = \mathrm{proj}\left(Y \middle| \mathcal{H}_t^X\right)\ , \tag{2.16}$$

where the limit holds in the L^2 sense.

Definition 2.6.2 (Prediction coefficients)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. We call the* predictor of order p *the random variable* $\mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}^X\right)$ *and the* partial innovation process of order p *the process $\epsilon_p^+ = \left(\epsilon_{t,p}^+\right)_{t\in\mathbb{Z}}$ defined by*

$$\epsilon_{t,p}^+ = X_t - \mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}^X\right)\ .$$

The prediction coefficients *are any coefficients* $\phi_p^+ = \left(\phi_{k,p}^+\right)_{k=1,\dots,p}$ *which satisfy, for all* $t \in \mathbb{Z}$,

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}^X\right) = \sum_{k=1}^{p} \phi_{k,p}^+ X_{t-k} \ . \tag{2.17}$$

Observe that, by the orthogonality principle, (2.17) is equivalent to

$$\Gamma_p^+ \phi_p^+ = \gamma_p^+ \ , \tag{2.18}$$

where $\gamma_p^+ = [\gamma(1), \gamma(2), \cdots, \gamma(p)]^T$ and

$$\begin{aligned}\Gamma_p^+ &= \operatorname{Cov}\left([X_{t-1} \ \dots \ X_{t-p}]^T\right)^T \\ &= \begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & & \gamma(-p+1) \\ \gamma(1) & \gamma(0) & \gamma(-1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(-1) \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix} ,\end{aligned}$$

Observing that Equation (2.18) does not depend on t and that the orthogonal projection is always well defined, such coefficients $\left(\phi_{k,p}^+\right)_{k=1,\dots,p}$ always exist. However they are uniquely defined if and only if Γ_p^+ is invertible.

Let us now compute the variance of the order-p prediction error $\epsilon_{t,p}^+$, denoted as

$$\sigma_p^2 = \| X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \|^2 = \mathbb{E}\left[\left| X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \right|^2 \right] \ . \tag{2.19}$$

By (2.17) and Proposition A.4.2, we have

$$\begin{aligned}\sigma_p^2 &= \langle X_t, X_t - \operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}\right) \rangle \\ &= \gamma(0) - \sum_{k=1}^{p} \overline{\phi_{k,p}^+} \gamma(k) \\ &= \gamma(0) - (\phi_p^+)^H \gamma_p^+ \ . \end{aligned} \tag{2.20}$$

Equations (2.18) and (2.20) are called *Yule-Walker equations*. An important consequence of these equations is that σ_p^2 does not depend on t, and since (2.16) implies

$$\sigma^2 = \lim_{p \to \infty} \sigma_p^2 \ ,$$

we obtain that, as claimed above, the variance of the innovation defined in (2.15) is also independent of t. So we can state the following result.

Corollary 2.6.1. *The innovation process of a centered weakly stationary process* X *is a (centered) weak white noise. Its variance is called the* innovation variance of the process X.

The innovation variance is not necessarily positive, that is, the innovation process can be zero a.s., as shown by the following example.

Example 2.6.3 (Innovations of the harmonic process (continued from Example 2.3.2))**.** *Consider the harmonic process $X_t = A\cos(\lambda_0 t + \Phi)$ where A is a centered random variable with finite variance σ_A^2 and Φ is a random variable, independent of A, with uniform distribution on $(0, 2\pi)$. Then X is a centered weakly stationary process with autocovariance function $\gamma(\tau) = (\sigma_A^2/2)\cos(\lambda_0\tau)$. The prediction coefficients or order 2 are given by*

$$\begin{bmatrix}\phi_{1,2}^+\\ \phi_{2,2}^+\end{bmatrix} = \begin{bmatrix}1 & \cos(\lambda_0)\\ \cos(\lambda_0) & 1\end{bmatrix}^{-1}\begin{bmatrix}\cos(\lambda_0)\\ \cos(2\lambda_0)\end{bmatrix} = \begin{bmatrix}2\cos(\lambda_0)\\ -1\end{bmatrix}$$

We then obtain that $\sigma_2^2 = \|X_t - \mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,2}^X\right)\|^2 = 0$ and thus

$$X_t = \mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,2}^X\right) = 2\cos(\lambda_0)X_{t-1} - X_{t-2} \in \mathcal{H}_{t-1}^X$$

Hence in this case the innovation process is zero: one can exactly predict the value of X_t from its past.

The latter example indicates that the harmonic process is *deterministic*, according to the following definition.

Definition 2.6.3 (Regular/deterministic process)**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. If the variance of its innovation process is zero, we say that X is* deterministic*. Otherwise, we say that X is* regular*.*

Let us define the intersection of the whole past of the process X as

$$\mathcal{H}_{-\infty}^X = \bigcap_{t\in\mathbb{Z}} \mathcal{H}_t^X \ .$$

Note that this (closed) linear space may not be null. Take a deterministic process X such as the harmonic process above. Then $X_t \in \mathcal{H}_{t-1}^X$, which implies that $\mathcal{H}_t^X = \mathcal{H}_{t-1}^X$. Thus, for a deterministic process, we have, for all t, $\mathcal{H}_{-\infty}^X = \mathcal{H}_t^X$, and thus also, $\mathcal{H}_{-\infty}^X = \mathcal{H}^X$, which is of course never null unless $X = 0$ a.s.

Example 2.6.4 (Constant process)**.** *A very simple example of deterministic process is obtained by taking $\lambda_0 = 0$ in Example 2.6.3. In other words, $X_t = X_0$ for all $t \in \mathbb{Z}$.*

For a regular process, things are a little bit more involved. For the white noise, it is clear that $\mathcal{H}_{-\infty}^X = \{0\}$. In this case, we say that X is *purely non-deterministic*. However not every regular process is purely nondeterministic. Observe indeed that for two uncorrelated centered and weakly stationary process X and Y, setting $Z = X + Y$, which is also centered and weakly stationary, we have, for all $t \in \mathbb{Z}$

$$\mathcal{H}_t^Z \subseteq \mathcal{H}_t^X \overset{\perp}{\oplus} \mathcal{H}_t^Y \ .$$

This implies that

$$\mathcal{H}_{-\infty}^Z \subseteq \mathcal{H}_{-\infty}^X \overset{\perp}{\oplus} \mathcal{H}_{-\infty}^Y \ . \tag{2.21}$$

Also, by Proposition A.4.2, the innovation variance of Z is larger than the sum of the innovations variances of X and Y. From these facts, we have that the sum of two uncorrelated processes is regular if at least one of them is regular and it is purely non-deterministic if both are purely non-deterministic. A regular process which is not purely nondeterministic can easily be obtained as follows.

Example 2.6.5 (Uncorrelated sum of a white noise with a constant process)**.** *Define $Z = X + Y$ with $X \sim \mathrm{WN}(0, \sigma^2)$ and $Y_t = Y_0$ for all t, where Y_0 is centered with positive variance and uncorrelated with $(X_t)_{t\in\mathbb{Z}}$. Then by (2.21), $\mathcal{H}^Z_{-\infty} \subseteq \mathrm{Span}\,(Y_0)$. Moreover, it can be shown (see Exercise 2.10) that $Y_0 \in \mathcal{H}^Z_{-\infty}$ and thus $X_t = Z_t - Y_0 \in \mathcal{H}^Z_t$. Hence we obtain $\mathcal{H}^Z_{-\infty} = \mathrm{Span}\,(Y_0)$, so that Z is not purely non-deterministic and Z has innovation X, so that Z is regular.*

In fact, the Wold decomposition indicates that the configuration of Example 2.6.5 is the only one: every regular process is the sum of two uncorrelated processes: one which is deterministic, the other which is purely nondeterministic. Before stating this result we introduce the following coefficients, defined for any regular process X,

$$\psi_s = \frac{\langle X_t, \epsilon_{t-s}\rangle}{\sigma^2}\ , \tag{2.22}$$

where ϵ is the innovation process and σ^2 its variance. By weak stationarity of X, this coefficient do no depend on t but only on k, since

$$\begin{aligned}\langle X_t, \epsilon_{t-k}\rangle &= \gamma(k) - \mathrm{Cov}\left(X_t, \mathrm{proj}\left(X_{t-k}|\,\mathcal{H}^X_{t-k-1}\right)\right)\\ &= \gamma(k) - \lim_{p\to\infty} \mathrm{Cov}\left(X_t, \mathrm{proj}\left(X_{t-k}|\,\mathcal{H}^X_{t-k-1,p}\right)\right)\\ &= \gamma(k) - \lim_{p\to\infty} \sum_{j=1}^{p} \phi_{j,p}\gamma(k+j)\ .\end{aligned}$$

It is easy to show that $\psi_0 = 1$. Moreover, since ϵ is a white noise, we have, for all $t \in \mathbb{Z}$,

$$\mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t) = \sum_{k\geq 0} \psi_k \epsilon_{t-k}\ .$$

We can now state the Wold decomposition.

Theorem 2.6.2 (Wold decomposition)**.** *Let X be a regular process and let ϵ be its innovation process and σ^2 its innovation variance, so that $\epsilon \sim \mathrm{WN}(0, \sigma^2)$. Define the L^2 centered process U as*

$$U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}\ ,$$

where ψ_k is defined by (2.22). Defined the L^2 centered process V by the following equation:

$$X_t = U_t + V_t, \quad \textit{for all } t \in \mathbb{Z}. \tag{2.23}$$

Then the following assertions hold.

(i) We have $U_t = \mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t)$ and $V_t = \mathrm{proj}\left(X_t|\,\mathcal{H}^X_{-\infty}\right)$.

(ii) ϵ and V are uncorrelated: for all (t, s), $\langle V_t, \epsilon_s\rangle = 0$.

(iii) U is a purely non-deterministic process and has same innovation as X. Moreover, $\mathcal{H}^\epsilon_t = \mathcal{H}^U_t$ for all $t \in \mathbb{Z}$.

(iv) V is a deterministic process and $\mathcal{H}^V_{-\infty} = \mathcal{H}^X_{-\infty}$.

Proof. The proof mainly relies on the facts established above and on Theorem A.4.3. Notice that, for all $s \leq t$ we have

$$\mathcal{H}_s^X \overset{\perp}{\oplus} \operatorname{Span}(\epsilon_k,\, s < k \leq t) = \mathcal{H}_t^X \ .$$

Then, by Theorem A.4.3, we get that

$$\mathcal{H}_{-\infty}^X \overset{\perp}{\oplus} \mathcal{H}_t^\epsilon = \mathcal{H}_t^X \ .$$

The facts then easily follow. The details are left to the reader (see Exercise 2.11). □

2.7 Linear forecasting of a weakly stationary time series

2.7.1 Choleski decomposition

Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with autocovariance function γ. We already have considered the problem of p-th order linear prediction of X_t by a *linear predictor* defined as a linear combination of $X_{t-1}, \ldots, X_{t-p}$. The optimal coefficients are called the linear predictor coefficients, see Definition 2.6.2. More precisely, they are defined as $\phi_p^+ = \begin{bmatrix}\phi_{1,p}^+ & \ldots & \phi_{p,p}^+\end{bmatrix}^T$ with

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}_{t-1,p}^X\right) = \sum_{k=1}^{p} \phi_{k,p}^+ X_{t-k} \ ,$$

which is equivalent to

$$\Gamma_p^+ \phi_p^+ = \gamma_p^+ \ , \tag{2.24}$$

where $\gamma_p^+ = \begin{bmatrix}\gamma(1) & \gamma(2) & \cdots & \gamma(p)\end{bmatrix}^T$ and

$$\begin{aligned}\Gamma_p^+ &= \operatorname{Cov}\left(\begin{bmatrix}X_{t-1} & \ldots & X_{t-p}\end{bmatrix}^T\right)^T \\ &= \begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & & \gamma(-p+1) \\ \gamma(1) & \gamma(0) & \gamma(-1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(-1) \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix} ,\end{aligned}$$

We are now interested in the effective computation of the prediction coefficients ϕ_p^+ (given γ) and of the prediction error defined by (2.19) and given by

$$\sigma_p^2 = \gamma(0) - (\phi_p^+)^H \gamma_p^+ \ , \tag{2.25}$$

see (2.20). The equations (2.24) and (2.25) are generally referred to as the *Yule-Walker equations.*

Obviously the Yule-Walker equations have a unique solution (ϕ_p^+, σ_p^2) if and only if Γ_p^+ is invertible. Proposition 2.4.3 provides a very simple (and general) sufficient condition for the invertibility of Γ_p^+, namely if $\gamma(0) \neq 0$ and $\gamma(t) \to 0$ as $t \to \infty$.

The following theorem induces a more precise condition. It also provides a Choleski decomposition of Γ_p^+.

Theorem 2.7.1. *Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with autocovariance function γ. Let $\sigma_0^2 = \gamma(0)$ and for all $p \geq 1$, (ϕ_p^+, σ_p^2) be any solution of the Yule-Walker equations (2.24) and (2.25). Then we have, for all $p = 0, 1, \ldots,$*

$$\Gamma_{p+1}^+ = A_{p+1}^{-1} D_{p+1} (A_{p+1}^H)^{-1} , \tag{2.26}$$

where

$$A_{p+1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\phi_{1,1}^+ & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ -\phi_{p,p}^+ & -\phi_{p-1,p}^+ & \cdots & -\phi_{1,p}^+ & 1 \end{bmatrix} ,$$

and

$$D_{p+1} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & & \cdots & \sigma_p^2 \end{bmatrix} .$$

In particular, Γ_{p+1}^+ is invertible if and only if $\sigma_p^2 > 0$ and, if X is a regular process, then Γ_p^+ is invertible for all $p \geq 1$.

Proof. Denote

$$\mathbf{X}_{p+1} = \begin{bmatrix} X_1 & \ldots & X_{p+1} \end{bmatrix}^T .$$

By Definition 2.6.2, we have

$$\begin{aligned} A_{p+1}\mathbf{X}_{p+1} &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\phi_{1,1}^+ & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ -\phi_{p,p}^+ & -\phi_{p-1,p}^+ & \cdots & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{p+1} \end{bmatrix} \\ &= \begin{bmatrix} X_1 \\ X_2 - \operatorname{proj}\left(X_2 \middle| \mathcal{H}_{1,1}^X\right) \\ \vdots \\ X_{p+1} - \operatorname{proj}\left(X_{p+1} \middle| \mathcal{H}_{p,p}^X\right) \end{bmatrix} \\ &= \begin{bmatrix} X_1 \\ \epsilon_{2,1}^+ \\ \vdots \\ \epsilon_{p+1,p}^+ \end{bmatrix} . \end{aligned}$$

Observe that, for all $k \geq 1$, $\mathcal{H}_{k,k}^X = \operatorname{Span}(X_1, \ldots, X_k)$ and thus increases with k. Using that $X_1 \in \mathcal{H}_{1,1}^X$ and for all $k = 2, \ldots, p$, $\epsilon_{k,k-1}^+ \in \mathcal{H}_{k,k}^X$ and $\epsilon_{k+1,k}^+ \perp \mathcal{H}_{k,k}^X$, we get that $\begin{bmatrix} X_1 & \epsilon_{2,1}^+ & \ldots & \epsilon_{p+1,p}^+ \end{bmatrix}^T$ have orthogonal components with variances $\sigma_0^2, \ldots, \sigma_p^2$. Hence we obtain

$$\operatorname{Cov}(A_{p+1}\mathbf{X}_{p+1}) = D_{p+1} ,$$

from which we get (2.26). □

It is interesting to observe that the prediction coefficients can be defined using the spectral measure ν of X. Indeed by definition of the orthogonal projection we have

$$\phi_p^+ = \underset{\phi\in\mathbb{C}^p}{\operatorname{argmin}}\, \mathbb{E}\left[X_t - \begin{bmatrix} X_{t-1} & \dots & X_{t-p}\end{bmatrix}\phi\right] .$$

and

$$\sigma_p^2 = \inf_{\phi\in\mathbb{C}^p} \mathbb{E}\left[X_t - \begin{bmatrix} X_{t-1} & \dots & X_{t-p}\end{bmatrix}\phi\right] .$$

Now for all $\phi \in \mathbb{Z}$, we have

$$\mathbb{E}\left[\left|X_t - \begin{bmatrix} X_{t-1} & \dots & X_{t-p}\end{bmatrix}\phi\right|^2\right] = \int_{\mathbb{T}} |\Phi(\mathrm{e}^{-\mathrm{i}\lambda})|^2 \,\mathrm{d}\nu(\lambda) ,$$

where Φ is the polynomial defined by

$$\Phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k .$$

Using this approach, the following interesting result can be shown. The detailed proof is left to the reader (see Exercise 2.12).

Theorem 2.7.2. *Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with autocovariance function γ. Let $\sigma_0^2 = \gamma(0)$ and for all $p \geq 1$, (ϕ_p^+, σ_p^2) be any solution of the Yule-Walker equations (2.24) and (2.25). Then, if Γ_p^+ is invertible we have for all z in the closed unit disk $\{z \in \mathbb{C}, |z| \leq 1\}$,*

$$1 - \sum_{k=1}^{p} \phi_{k,p}^+ z^k \neq 0 .$$

2.7.2 Levinson-Durbin Algorithm

The usual way to compute the inverse of a symmetric positive definite matrix is to rely on the Choleski decomposition, which requires $O(p^3)$ operations. However this approach does not take advantage of the particular geometric structure of the matrices Γ_p^+. We now introduce a more efficient recursive algorithm that allows to solve the Yule-Walker equations in $O(p^2)$ operations.

Algorithm 1: Levinson-Durbin algorithm.

Data: Covariance coefficients $\gamma(k)$, $k = 0, \dots, K$
Result: Prediction coefficients $\{\phi^+_{m,p}\}_{1\le m\le p, 1\le p\le K}$, partial autocorrelation coefficients $\kappa(1), \dots, \kappa(K)$
Initialization: set $\kappa(1) = \gamma(1)/\gamma(0)$, $\phi^+_{1,1} = \gamma(1)/\gamma(0)$, $\sigma_1^2 = \gamma(0)(1 - |\kappa(1)|^2)$.
for $p = 1, 2, \dots, K-1$ **do**
  Set

$$\kappa(p+1) = \sigma_p^{-2}\left(\gamma(p+1) - \sum_{k=1}^{p} \phi^+_{k,p}\gamma(p+1-k)\right) \tag{2.27}$$

$$\sigma^2_{p+1} = \sigma_p^2(1 - |\kappa(p+1)|^2) \tag{2.28}$$

$$\phi^+_{p+1,p+1} = \kappa(p+1) \tag{2.29}$$

  for $m \in \{1, \cdots, p\}$ **do**
    Set

$$\phi^+_{m,p+1} = \phi^+_{m,p} - \kappa(p+1)\,\overline{\phi^+_{p+1-m,p}}\,. \tag{2.30}$$

  end
end

Observe that all the computations of Algorithm 1 can be done in $O(K^2)$ operations.

Theorem 2.7.3. *Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with autocovariance function γ. Let $\sigma_0^2 = \gamma(0)$ and for all $p \ge 1$, (ϕ_p^+, σ_p^2) be any solution of the Yule-Walker equations (2.24) and (2.25). Then Algorithm 1 applies for any K such that Γ_K^+ is invertible, or, equivalently, $\sigma^2_{K-1} > 0$.*

Before proving this theorem, let us state an important and useful lemma.

Lemma 2.7.4. *Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with autocovariance function γ. Let $\epsilon^+_{t,0} = \epsilon^-_{t,0} = X_t$ and, for $p \ge 1$, $\epsilon^+_{t,p}$ and $\kappa(p)$ are as in Definition 2.6.2 and Definition 3.5.1. Define moreover the* backward partial innovation process of order $p \ge 1$ *by*

$$\epsilon^-_{t,p} = X_t - \mathrm{proj}\left(X_t \middle| \mathcal{H}^X_{t+p,p}\right)\,.$$

Then, for all $p \ge 0$, we have $\|\epsilon^+_{t,p}\| = \|\epsilon^-_{t-p-1,p}\|$ and

$$\kappa(p+1) = \frac{\left\langle \epsilon^+_{t,p}, \epsilon^-_{t-p-1,p}\right\rangle}{\|\epsilon^+_{t,p}\|\ \|\epsilon^-_{t-p-1,p}\|}\,, \tag{2.31}$$

with the convention $0/0 = 0$.

Proof. Let us denote by c the right-hand side of (2.31) in this proof, that is,

$$c = \frac{\left\langle \epsilon^+_{t,p}, \epsilon^-_{t-p-1,p}\right\rangle}{\|\epsilon^+_{t,p}\|\ \|\epsilon^-_{t-p-1,p}\|}\,.$$

The result is straightforward for $p = 0$ since in this case $\epsilon^+_{t,p} = X_t$ and $\epsilon^-_{t-p-1,p} = X_{t-1}$.

We now take $p \geq 1$. Observe that

$$\begin{aligned}\|\epsilon_{t,p}^{+}\|^2 &= \inf_{Y \in \mathcal{H}_{t-1,p}^{X}} \|X_t - Y\|^2 \\ &= \inf_{\phi \in \mathbb{C}^p} \begin{bmatrix} 1 & -\phi^T \end{bmatrix} \Gamma_{p+1}^{+} \begin{bmatrix} 1 & -\phi^T \end{bmatrix}^H \ ,\end{aligned}$$

where we used that $\Gamma_{p+1}^{+} = \mathrm{Cov}\left(\begin{bmatrix} X_t & X_{t-1} & \dots & X_{t-p} \end{bmatrix}^T\right)$. Similarly, we have

$$\begin{aligned}\|\epsilon_{t,p}^{-}\|^2 &= \inf_{Y \in \mathcal{H}_{t+p,p}^{X}} \|X_t - Y\|^2 \\ &= \inf_{\phi \in \mathbb{C}^p} \begin{bmatrix} 1 & -\phi^T \end{bmatrix} \Gamma_{p+1}^{-} \begin{bmatrix} 1 & -\phi^T \end{bmatrix}^H \ ,\end{aligned}$$

where we used that $\Gamma_{p+1}^{-} = \mathrm{Cov}\left(\begin{bmatrix} X_t & X_{t+1} & \dots & X_{t+p} \end{bmatrix}^T\right)$. Using that γ is hermitian we get

$$\Gamma_{p+1}^{-} = \overline{\Gamma_{p+1}^{+}} \ .$$

Hence we have

$$\sigma_p^2 = \|\epsilon_{t,p}^{+}\|^2 = \|\epsilon_{t,p}^{-}\|^2 \ .$$

Observe that $\epsilon_{t-p-1,p}^{-} = X_{t-p-1} - \mathrm{proj}\left(X_{t-p-1} | \, \mathcal{H}_{t-1,p}^{X}\right)$. Hence,

$$c = \frac{\left\langle \epsilon_{t,p}^{+}, \epsilon_{t-p-1,p}^{-} \right\rangle}{\sigma_p^2} = \frac{\left\langle \epsilon_{t,p}^{+}, X_{t-p-1} \right\rangle}{\|\epsilon_{t-p-1,p}^{+}\|^2} = \frac{\left\langle X_t, \epsilon_{t-p-1,p}^{-} \right\rangle}{\|\epsilon_{t-p-1,p}^{-}\|^2} \ . \tag{2.32}$$

Moreover, we have

$$\begin{aligned}\mathcal{H}_{t-1,p+1}^{X} &= \mathrm{Span}\left(X_{t-1}, X_{t-1}, \dots, X_{t-1-p}\right) \\ &= \mathcal{H}_{t-1,p}^{X} + \mathrm{Span}\left(X_{t-p-1}\right) \\ &= \mathcal{H}_{t-1,p}^{X} \overset{\perp}{\oplus} \mathrm{Span}\left(\epsilon_{t-p-1,p}^{-}\right) \ .\end{aligned}$$

By Assertion (*vii*) in Proposition A.4.2, we get

$$\mathrm{proj}\left(X_t | \, \mathcal{H}_{t-1,p+1}^{X}\right) = \mathrm{proj}\left(X_t | \, \mathcal{H}_{t-1,p}^{X}\right) + \mathrm{proj}\left(X_t | \, \mathrm{Span}\left(\epsilon_{t-p-1,p}^{-}\right)\right) \ ,$$

and

$$\begin{aligned}&\|X_t - \mathrm{proj}\left(X_t | \, \mathcal{H}_{t-1,p+1}^{X}\right)\|^2 \\ &\qquad = \|X_t - \mathrm{proj}\left(X_t | \, \mathcal{H}_{t-1,p}^{X}\right)\|^2 - \left\|\mathrm{proj}\left(X_t | \, \mathrm{Span}\left(\epsilon_{t-p-1,p}^{-}\right)\right)\right\|^2 \ .\end{aligned} \tag{2.33}$$

Now we consider two cases.

First assume that $\sigma_p^2 \neq 0$. Then $\epsilon_{t-p-1,p}^{-}$ is non-zero, and we have, as in Example A.4.1

$$\mathrm{proj}\left(X_t | \, \mathrm{Span}\left(\epsilon_{t-p-1,p}^{-}\right)\right) = \frac{\left\langle X_t, \epsilon_{t-p-1,p}^{-} \right\rangle}{\|\epsilon_{t-p-1,p}^{-}\|^2} \, \epsilon_{t-p-1,p}^{-} = c \, \epsilon_{t-p-1,p}^{-} \ ,$$

where we used (2.32). Moreover, by Theorem 2.7.1, $\sigma_p^2 \neq 0$ implies that Γ_p^+ and Γ_{p+1}^+ are invertible, so ϕ_p^+ and ϕ_{p+1}^+ are uniquely defined by (2.24) and the last two displays give

$$\sum_{k=1}^{p+1} \phi_{k,p+1}^+ X_{t-k} = \sum_{k=1}^{p} \phi_{k,p}^+ X_{t-k} + c \left(X_{t-p-1} - \sum_{k=1}^{p} \phi_{k,p}^- X_{t-p-1+k} \right) , \tag{2.34}$$

where $\phi_p^- = \left(\phi_{k,p}^-\right)_{k=1,\dots,p}$ is uniquely defined by

$$\operatorname{proj}\left(X_{t-p-1} \middle| \mathcal{H}_{t-1,p}^X \right) = \sum_{k=1}^{p} \phi_{k,p}^- X_{t-p-1+k} \, . \tag{2.35}$$

Since the prediction coefficients are uniquely defined in (2.34), we get by identifying those of the left-hand side with those of the right-hand side that

$$\phi_{k,p+1}^+ = \phi_{k,p}^+ - c\phi_{p+1-k,p}^- \qquad \text{for } k = 1, \dots, p \tag{2.36}$$

$$\phi_{p+1,p+1}^+ = c \tag{2.37}$$

Equation (2.37) gives (2.31), which concludes the proof in the case where $\sigma_p^2 \neq 0$.

In the case where $\sigma_p^2 = 0$, then, by convention $c = 0$. By Theorem 2.7.1, we also have that Γ_{p+1}^+ is not invertible so that $\kappa(p+1) = 0$ by the convention in Definition 3.5.1. □

The proof of Theorem 2.7.3 can now be completed.

Proof of Theorem 2.7.3. The initialization step is straightforward, see Example A.4.1.

We now prove the iteration formula, that is (2.27), (2.29), (2.28) and (2.30). Relation (2.29) is proved in Lemma 2.7.4. Under the assumptions of Theorem 2.7.3, we can use the facts shown in the proof of Lemma 2.7.4 in the case where $\sigma_p^2 \neq 0$. Relation (2.32) gives that

$$\kappa(p+1) = \frac{\left\langle X_t - \phi_p^{+T} \begin{bmatrix} X_{t-1} & \dots & X_{t-p} \end{bmatrix}, X_{t-p-1} \right\rangle}{\sigma_p^2} ,$$

which yields (2.27).

Relation (2.33) implies that

$$\sigma_{p+1}^2 = \sigma_p^2 - |c|^2 \sigma_p^2 ,$$

that is, by definition of c, we get (2.28).

To prove (2.30) with (2.36), we need to relate ϕ_p^- (uniquely defined by (2.35) with ϕ_p^+, solution of (2.24). By Theorem A.4.1, ϕ_p^- is the unique solution of

$$\Gamma_p^- \phi_p^- = \gamma_p^-$$

where $\gamma_p^- = [\gamma(-1), \gamma(-2), \cdots, \gamma(-p)]^T$ and

$$\Gamma_p^- = \operatorname{Cov}\left([X_{t-p} \ \dots \ X_{t-1}]^T \right)^T \tag{2.38}$$

$$= \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & & \gamma(p-1) \\ \gamma(-1) & \gamma(0) & \gamma(1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(1) \\ \gamma(1-p) & \gamma(2-p) & \cdots & \gamma(-1) & \gamma(0) \end{bmatrix} . \tag{2.39}$$

Hence $\gamma_p^- = \overline{\gamma_p^+}$ and $\Gamma_p^- = \overline{\Gamma_p^+}$ and so $\phi_p^- = \overline{\phi_p^+}$. This, with (2.36), yields (2.30), which concludes the proof. □

2.7.3 The innovations algorithm

The Levinson-Durbin algorithm provides the prediction coefficients and prediction error variances, and thus also the Choleski decomposition of Γ_p^+, see Theorem 2.7.1. In contrast, the innovation algorithm allows us to iteratively compute the predictors of finite order and the prediction errors variances by expressing the predictors in an orthogonal basis, rather than the original time series. It is in fact the Gram-Schmidt procedure (see Algorithm 12) applied in our particular context. A significant advantage of the innovation algorithm is that it also applies if X is non-stationary.

To deal with non-stationary time series, we adapt the definitions of innovations. We consider in this section a centered L^2 process $(X_t)_{n\in\mathbb{N}}\, t \geq 1$ with covariance function

$$\gamma(j,k) = \mathrm{Cov}\left(X_j, X_k\right), \quad j,k \geq 1\ . \tag{2.40}$$

Further define $\mathcal{H}_q^X = \mathrm{Span}\left(X_1, \ldots, X_q\right)$ and the innovation process

$$\epsilon_1 = X_1 \quad \text{and} \quad \epsilon_t = X_t - \mathrm{proj}\left(X_t | \, \mathcal{H}_{t-1}^X\right), \quad t = 2, 3, \ldots \tag{2.41}$$

As in Example A.2.1, we immediately obtain that $(\epsilon_t)_{n\in\mathbb{N}}\, t \geq 1$ is an orthogonal sequence, moreover we have, for all $p \geq 1$,

$$\mathcal{H}_p^X = \mathrm{Span}\left(\epsilon_1, \ldots, \epsilon_p\right)\ .$$

We denote by $\theta_p = (\theta_{k,p})_{k=1,\ldots,p}$ the coefficients of the linear predictor $\mathrm{proj}\left(X_{p+1} | \, \mathcal{H}_p^X\right)$ in this basis,

$$\mathrm{proj}\left(X_{p+1} | \, \mathcal{H}_p^X\right) = \sum_{k=1}^{p} \theta_{k,p}\epsilon_k\ .$$

and by σ_p^2 the prediction error variance

$$\sigma_p^2 = \left\| X_{p+1} - \mathrm{proj}\left(X_{p+1} | \, \mathcal{H}_p^X\right)\right\|^2 = \|\epsilon_{p+1}\|^2\ .$$

In this context the following algorithm applies.

Algorithm 2: Innovation algorithm.

Data: Covariance coefficients $\gamma(k,j)$, $1 \le j \le k \le K+1$, observed variables $X_1, \dots, X_{K+1}$

Result: Innovation variables $\epsilon_1, \dots, \epsilon_{K+1}$, prediction coefficients $\theta_p = (\theta_{k,p})_{k=1,\dots,p}$ in the innovation basis and prediction error variances σ_p^2 for $p = 1, \dots, K$.

Initialization: set $\sigma_0^2 = \gamma(1,1)$ and $\epsilon_1 = X_1$.

for $p = 1, \dots, K$ **do**

  for $m = 1, \dots, p$ **do**

    Set

$$\theta_{m,p} = \sigma_{m-1}^{-2} \left(\gamma(p+1, m) - \sum_{j=1}^{m-1} \overline{\theta_{j,m-1}}\, \theta_{j,p}\, \sigma_j^2 \right)$$

  end

  Set

$$\sigma_p^2 = \gamma(p+1, p+1) - \sum_{m=1}^{p} |\theta_{m,p}|^2\, \sigma_{m-1}^2$$

$$\epsilon_{p+1} = X_{p+1} - \sum_{m=1}^{p} \theta_{m,p} \epsilon_m \, .$$

end

Of course Algorithm 2 applies also in the case where X is weakly stationary. Observe that all the computations of Section 2.7.3 can be done in $O(K^3)$ operations. Hence in the weakly stationary case, one should prefer Algorithm 1 to Algorithm 2. On the other hand, there is one case where Algorithm 2 can be achieved in $O(K)$ operations, namely, if X is an MA(q) process, since in this case,

$$t > s + q \Rightarrow X_t \perp \mathcal{H}_s^X \, ,$$

and thus we have

$$\theta_{k,p} = 0 \quad \text{for all } k < p + 1 - q$$

A particular application is examined in the following example.

Example 2.7.1 (Prediction of an MA(1) process)**.** *Let $X_t = Z_t + \theta Z_{t-1}$ where $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ and $\theta \in \mathbb{C}$. It follows that $\gamma(i,j) = 0$ for all $|i-j| > 1$, $\gamma(i,i) = \sigma^2(1+|\theta|^2)$ et $\gamma(i+1,i) = \theta\sigma^2$. Moreover Algorithm 2 boils down to*

$$\begin{aligned}
&\sigma_0^2 = (1+|\theta|^2)\sigma^2 \, , \\
&\sigma_p^2 = \sigma^2 \left(1 + |\theta|^2 - \sigma_{p-1}^{-2} |\theta|^2 \sigma^2\right) , && p \ge 1 \, , \\
&\theta_{k,p} = 0 \, , && 1 \le k \le p-1 \, , \\
&\theta_{p,p} = \sigma_{p-1}^{-2} \theta \sigma^2 \, , && p \ge 1 \, .
\end{aligned}$$

Setting $r_p = \sigma_p^2/\sigma^2$, we get

$$\begin{aligned}
&\epsilon_1 = X_1 \, , \\
&\epsilon_{p+1} = X_{p+1} - \theta \epsilon_p / r_{n-1} \, , && p \ge 1 \, ,
\end{aligned}$$

with $r_0 = 1 + \theta^2$, and for $p \ge 1$, $r_{p+1} = 1 + \theta^2 - \theta^2/r_p$.

2.8 Exercises

Exercise 2.1. Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. real valued random variables. Determine in each of the following cases, if the defined process is strongly stationary.

1. $Y_t = a + b\varepsilon_t + c\varepsilon_{t-1}$ (a, b, c real numbers).
2. $Y_t = a + b\varepsilon_t + c\varepsilon_{t+1}$.
3. $Y_t = \sum_{j=0}^{+\infty} \rho^j \varepsilon_{t-j}$ for $|\rho| < 1$, assuming that $\mathbb{E}\left[|\epsilon_0|\right] < \infty$.
4. $Y_t = \varepsilon_t \varepsilon_{t-1}$.
5. $Y_t = (-1)^t \varepsilon_t$, $Z_t = \varepsilon_t + Y_t$.

Exercise 2.2. Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ be two second order stationary processes that are uncorrelated in the sense that X_t and Y_s are uncorrelated for all t, s. Show that $Z_t = X_t + Y_t$ is a second order stationary process. Compute its autocovariance function, given the autocovariance functions of X and Y. Do the same for the spectral measures.

Exercise 2.3. Consider the processes of Exercise 2.1, with the additional assumption that $(\varepsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$. Determine in each case, if the defined process is weakly stationary. In the case of Question 4, consider also $Z_t = Y_t^2$ under the assumption $\mathbb{E}[\varepsilon_0^4] < \infty$.

Exercise 2.4. Define χ as in (2.7).

1. For which values of ρ is χ an autocovariance function ? [Hint : use the Hertglotz theorem].
2. Exhibit a Gaussian process with autocovariance function χ.

Exercise 2.5. For $t \geq 2$, define

$$\Sigma_2 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \ldots, \Sigma_t = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$$

1. For which values of ρ, is Σ_t guaranteed to be a covariance matrix for all values of t [Hint: write Σ_t as $\alpha I + A$ where A has a simple eigenvalue decomposition]?
2. Define a stationary process whose finite-dimensional covariance matrices coincide with Σ_t (for all $t \geq 1$).

Exercise 2.6. Let X and Y two L^2 centered random variables. Define

$$\rho = \frac{\mathrm{Cov}\,(X, Y)}{\mathrm{Var}\,(Y)}\,,$$

with the convention $0/0 = 0$. Show that

$$\mathrm{proj}\,(X|\,\mathrm{Span}\,(Y)) = \rho\, Y \quad \text{and} \quad \mathbb{E}\left[(X - \mathrm{proj}\,(X|\,\mathrm{Span}\,(Y)))^2\right] = \mathrm{Var}\,(X) - |\rho|^2\,\mathrm{Var}\,(Y)\ .$$

Exercise 2.7. Let (Y_t) be a weakly stationary process with spectral density function f such that $0 \leq m \leq f(\lambda) \leq M < \infty$ for all $\lambda \in \mathbb{R}$. For $n \geq 1$, denote by Γ_n the covariance matrix of $[Y_1, \ldots, Y_n]^T$. Show that the eigenvalues of Γ_n belong to the interval $[2\pi m, 2\pi M]$.

Exercise 2.8. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with spectral density function f and denote by $\hat{X}$ its spectral representation random measure, so that, for all $t \in \mathbb{Z}$,

$$X_t = \int \mathrm{e}^{\mathrm{i}t\lambda} \, \mathrm{d}\hat{X}(\lambda) \, .$$

Assume that f is two times continuously differentiable and that $f(0) = 0$. Define, for all $t \geq 0$,

$$Y_t = X_{-t} + X_{-t+1} + \cdots + X_0 \, .$$

1. Build an example of such a process X of the form $X_t = \epsilon_t + a\epsilon_{t-1}$ with $\epsilon \sim \mathrm{WN}(0, 1)$ and $a \in \mathbb{R}$.

2. Determine g_t such that $Y_t = \int g_t \, \mathrm{d}\hat{X}$.

3. Compute

$$\lim_{n\to\infty} \int_{\mathbb{T}} \left| \frac{1}{n} \sum_{k=1}^{n} \mathrm{e}^{-\mathrm{i}k\lambda} \right|^2 \mathrm{d}\lambda$$

4. Show that

$$Z = \int (1 - \mathrm{e}^{-\mathrm{i}\lambda})^{-1} \, \mathrm{d}\hat{X}(\lambda) \, .$$

   is well defined in $\mathcal{H}_\infty^X$.

5. Deduce from the previous questions that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} Y_t = Z \quad \text{in } L^2.$$

6. Show this result directly in the particular case exhibited in Question 1.

Exercise 2.9. Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with covariance function γ. Denote

$$\Gamma_t = \mathrm{Cov}\left(\left[X_1, \ldots, X_t\right]^T, = \right) \left[\gamma(i-j)\right]_{1\leq i,j\leq t} \qquad \text{for all } t \geq 1.$$

We temporarily assume that there exists $k \geq 1$ such that Γ_k is invertible but Γ_{k+1} is not.

1. Show that we can write X_n as $\sum_{t=1}^{k} \alpha_t^{(n)} X_t$, where $\alpha^{(n)} \in \mathbb{R}^k$, for all $n \geq k+1$.

2. Show that the vectors $\alpha^{(n)}$ are bounded independently of n.

Suppose now that $\gamma(0) > 0$ and $\gamma(t) \to 0$ as $t \to \infty$.

3. Show that, for all $t \geq 1$, Γ_t is invertible.

4. Deduce that Proposition 2.4.3 holds.

Exercise 2.10. Define $Z = X + Y$ with $X \sim \mathrm{WN}(0, \sigma^2)$ and $Y_t = Y_0$ for all t, where Y_0 is centered with positive variance and uncorrelated with $(X_t)_{t\in\mathbb{Z}}$.

1. Show that $\mathcal{H}^Z_{-\infty} \subseteq \mathrm{Span}\,(Y_0)$. [Hint : see Example 2.6.5]

Define, for all $t \in \mathbb{Z}$ and $n \geq 1$,

$$T_{t,n} = \frac{1}{n}\sum_{k=1}^{n} Z_{t-k}$$

2. What is the L^2 limit of $T_{t,n}$ as $n \to \infty$?
3. Deduce that $\mathcal{H}^Z_{-\infty} = \mathrm{Span}\,(Y_0)$.

Exercise 2.11. Define $(X_t)_{t\in\mathbb{Z}}$, $(U_t)_{t\in\mathbb{Z}}$ and $(V_t)_{t\in\mathbb{Z}}$ as in Theorem 2.6.2.

1. Show that
$$\mathcal{H}^X_{-\infty} \overset{\perp}{\oplus} \mathcal{H}^\epsilon_t = \mathcal{H}^X_t \ .$$
2. Deduce that $U_t = \mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t)$, $V_t = \mathrm{proj}\,\big(X_t|\,\mathcal{H}^X_{-\infty}\big)$ and that U and V are uncorrelated.
3. Show that $\mathcal{H}^X_{-\infty} = \mathcal{H}^V_t$ and $\mathcal{H}^\epsilon_t = \mathcal{H}^U_t$ for all $t \in \mathbb{Z}$. [Hint : observe that $\mathcal{H}^X_t \subset \mathcal{H}^U_t \oplus \mathcal{H}^V_t$ and use the previous questions]
4. Conclude the proof of Theorem 2.6.2.

Exercise 2.12. Let $(X_t)_{t\in\mathbb{Z}}$ and, (ϕ^+_p, σ^2_p), $p \geq 1$ be as in Theorem 2.7.2.

1. Compute (ϕ^+_1, σ^2_1) in the case where $\gamma(0) > 0$. Does $1 - \phi^+_{1,1}z$ vanish on the closed unit disk ?

Let $p \geq 2$ and suppose that Γ^+_p is invertible. Let ν^{-1} be a root of $\Phi(z) = 1 - \sum_{k=1}^p \phi^+_{k,p} z^k$, so that

$$\Phi(z) = (1 - \nu z)\Psi(z) \ ,$$

where Ψ is a polynomial of degree $p-1$. Define $Y = [\Psi(\mathrm{B})](X)$.

2. Show that
$$\mathbb{E}\left[(Y_1 - \nu Y_0)^2\right] = \inf_{\alpha\in\mathbb{C}} \mathbb{E}\left[(Y_1 - \alpha Y_0)^2\right]$$
Is ν uniquely defined by this equation ?
3. Conclude the proof of Theorem 2.7.2.

Chapter 3

ARMA models

In this chapter we focus on the linear filtering of time series. An important class of models for stationary time series, the autoregressive moving average (ARMA) models, are obtained by applying particular linear filters to a white noise. More general filters can be defined using the spectral representation of Section 2.5.3.

3.1 Linear filtering using absolutely summable coefficients

Let $\psi = (\psi_t)_{t\in\mathbb{Z}}$ be an absolutely summable sequence of $\mathbb{C}^{\mathbb{Z}}$, we will write $\psi \in \ell^1(\mathbb{Z})$, or simply $\psi \in \ell^1$.

In this section we consider the linear filter defined by

$$\mathrm{F}_\psi : x = (x_t)_{t\in\mathbb{Z}} \mapsto y = \psi \star x \;, \tag{3.1}$$

where $\star$ denotes the convolution product on sequences, that is, for all $t \in \mathbb{Z}$,

$$y_t = \sum_{k\in\mathbb{Z}} \psi_k x_{t-k} \;. \tag{3.2}$$

We introduce some usual terminology about such linear filters.

Definition 3.1.1. *We have the following definitions.*

(i) If ψ is finitely supported, F_ψ is called a finite impulse response (FIR) *filter.*

(ii) If $\psi_t = 0$ for all $t < 0$, F_ψ is said to be causal.

(iii) If $\psi_t = 0$ for all $t \geq 0$, F_ψ is said to be anticausal.

Of course (3.2) is not always well defined. In fact, F_ψ is well defined only on

$$\ell_\psi = \left\{ (x_t)_{t\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}} \;:\; \text{for all } t \in \mathbb{Z}, \sum_{k\in\mathbb{Z}} |\psi_k \, x_{t-k}| < \infty \right\} \;.$$

A natural question is to ask what happens for a random path, or in other words, given a random process $X = (X_t)_{t\in\mathbb{Z}}$, is $\mathrm{F}_\psi(X)$ well defined ? Observing that $\ell_\psi = \mathbb{C}^{\mathbb{Z}}$ if (and

only if) ψ has a finite support, this question is nontrivial only for an infinitely supported ψ. Moreover we observe that a FIR filter can be written as

$$\mathrm{F}_\psi = \sum_{k\in\mathbb{Z}} \psi_k \, \mathrm{B}^k \ , \tag{3.3}$$

where B is the Backshift operator of Definition 2.1.1. This sum is well defined for a finitely supported ψ since it is a finite sum of linear operators.

The following theorem provides an answer for $\psi \in \ell^1$ which always applies for a weakly stationary process X.

Theorem 3.1.1. *Let $\psi \in \ell^1$. Then, for all random process $X = (X_t)_{t\in\mathbb{Z}}$ such that*

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t| < \infty \ , \tag{3.4}$$

we have $X \in \ell_\psi$ a.s. If moreover

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left[|X_t|^2\right] < \infty \ , \tag{3.5}$$

then the series

$$Y_t = \sum_{k\in\mathbb{Z}} \psi_k X_{t-k} \ , \tag{3.6}$$

is absolutely convergent in L^2, and we have $(Y_t)_{t\in\mathbb{Z}} = \mathrm{F}_\psi(X)$ a.s.

Remark 3.1.1. *Recall that L^2 is complete, so an absolutely convergent series converges and $(Y_t)_{t\in\mathbb{Z}}$ is well defined and is an L^2 process.*

Proof of Theorem 3.1.1. We have, by the Tonelli theorem,

$$\mathbb{E}\left[\sum_{k\in\mathbb{Z}} |\psi_k X_{t-k}|\right] = \sum_{k\in\mathbb{Z}} |\psi_k| \, \mathbb{E}|X_{t-k}| \leq \sup_{t\in\mathbb{Z}} \mathbb{E}|X_t| \sum_{k\in\mathbb{Z}} |\psi_k| \ ,$$

which is finite by (3.4) and $\psi \in \ell^1$. Hence $X \in \ell_\psi$ a.s.

If (3.5) holds, the series in (3.6) is absolutely convergent in L^2 since

$$\sum_{k\in\mathbb{Z}} \left(\mathbb{E}\left[|\psi_k X_{t-k}|^2\right]\right)^{1/2} \leq \left(\sup_{t\in\mathbb{Z}} \mathbb{E}\left[|X_t|^2\right]\right)^{1/2} \sum_{k\in\mathbb{Z}} |\psi_k| < \infty \ ,$$

under Condition (3.5).

Finally, let us show that $(Y_t)_{t\in\mathbb{Z}}$ coincides with $\mathrm{F}_\psi(X)$ a.s. This follows from Fatou's Lemma. Denoting $\tilde{Y}_t = \Pi_t \circ \mathrm{F}_\psi(X)$ and

$$Y_{n,t} = \sum_{k=-n}^{n} \psi_k X_{t-k} \ ,$$

we get that

$$\mathbb{E}\left[|\tilde{Y}_t - Y_t|^2\right] = \mathbb{E}\left[\liminf_n |Y_{n,t} - Y_t|^2\right] \leq \liminf_n \mathbb{E}\left[|Y_{n,t} - Y_t|^2\right] = 0$$

which achieves the proof. □

An immediate consequence of this result is that F_ψ applies to any weakly stationary process and its output is also weakly stationary.

Theorem 3.1.2. *Let $\psi \in \ell^1$ and $X = (X_t)_{t\in\mathbb{Z}}$ be a weakly stationary process with mean μ, autocovariance function γ and spectral measure ν. Then $\mathrm{F}_\psi(X)$ is well defined and is a weakly stationary process with mean*

$$\mu' = \mu \sum_{t\in\mathbb{Z}} \psi_t \ , \tag{3.7}$$

autocovariance function given for all $h \in \mathbb{Z}$ by

$$\gamma'(h) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \psi_j \overline{\psi}_k \ \gamma_X(h+k-j) \ , \tag{3.8}$$

and spectral measure ν' defined as the measure with density $|\psi^*(\lambda)|^2$ with respect to ν, where

$$\psi^*(\lambda) = \sum_{t\in\mathbb{Z}} \psi_t \mathrm{e}^{-\mathrm{i}t\lambda} \ . \tag{3.9}$$

Proof. A weakly stationary processes satisfies the conditions of Theorem 3.1.1, hence $Y = \mathrm{F}_\psi(X)$ is well defined. Moreover Theorem 3.1.1 also says that each Y_t is obtained as the L^2 limit (3.6). By continuity and linearity of the mean in L^2, we get (3.7). Similarly, because the covariance defines a continuous inner product on L^2, we get (3.8).

Finally the spectral measure of Y is obtained by replacing γ in (3.8) by its spectral representation (see Theorem 2.4.1) and by the Fubini theorem (observing that ψ^* is bounded on $\mathbb{T}$). □

In the special case where X is a white noise, the above formulas simplify as follows.

Corollary 3.1.3. *Let $\psi \in \ell^1$ and $X \sim \mathrm{WN}(0, \sigma^2)$. Define $Y = \mathrm{F}_\psi(X)$. Then Y is a centered weakly stationary process with covariance function*

$$\gamma(h) = \sigma^2 \sum_{k\in\mathbb{Z}} \psi_{k+h} \overline{\psi}_k \ ,$$

and spectral density function

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{t\in\mathbb{Z}} \psi_t \mathrm{e}^{-\mathrm{i}t\lambda} \right|^2 \ .$$

one says that $Y = \mathrm{F}_\psi(X)$ is a centered linear process *with* short memory. *If moreover X is a strong white noise, then one says that $Y = \mathrm{F}_\psi(X)$ is a centered* strong linear process.

Here "short memory" refer to the fact that ψ is restricted to ℓ^1.

3.2 FIR filters inversion

Consider the following definition.

Definition 3.2.1. *Let $\psi \in \ell^1$ and X be a centered weakly stationary process. Let $Y = \mathrm{F}_\psi(X)$. We will say that this* linear representation *of Y is* invertible *if there exists $\phi \in \ell^1$ such that $X = \mathrm{F}_\phi(Y)$.*

This question of invertibility is of course very much related to the composition of filters. We have the following lemma.

Lemma 3.2.1. *Let $(\alpha_t)_{t\in\mathbb{Z}}$ and $(\beta_t)_{t\in\mathbb{Z}}$ be two sequences in ℓ^1. If X satisfies Condition (3.4), then*

$$\mathrm{F}_\alpha \circ \mathrm{F}_\beta(X) = \mathrm{F}_{\alpha\star\beta}(X) \quad a.s.$$

Proof. Denote $Y = \mathrm{F}_\beta(X)$. By Theorem 3.1.2, Y is well defined. Moreover, for all $t \in \mathsf{Z}$,

$$Y_t = \sum_{k\in\mathsf{Z}} \beta_k X_{t-k} \quad \text{a.s.} ,$$

so that

$$\mathbb{E}|Y_t| \leq \sup_{s\in\mathbb{Z}} \mathbb{E}|X_s| \times \sum_{k\in\mathsf{Z}} |\beta_k| < \infty .$$

Hence $\mathrm{F}_\alpha \circ \mathrm{F}_\beta$ is well defined on X a.s. and $Z = \mathrm{F}_\alpha \circ \mathrm{F}_\beta(X)$ satisfies, for all $t \in \mathsf{Z}$,

$$Z_t = \sum_{j\in\mathsf{Z}} \alpha_j Y_{t-j} \quad \text{a.s.} .$$

Observe also that $\alpha \star \beta \in \ell^1$ and define $W = \mathrm{F}_{\alpha\star\beta}(X)$. By Theorem 3.1.2, we have, for all $t \in \mathsf{Z}$, and

$$W_t = \sum_{k\in\mathsf{Z}} \left(\sum_{j\in\mathbb{Z}} \alpha_j \beta_{k-j} \right) X_{t-k} \quad \text{a.s.} .$$

Now by Tonelli's Theorem, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{k\in\mathsf{Z}} \sum_{j\in\mathbb{Z}} |\alpha_j \beta_{k-j} X_{t-k}| \right] &= \sum_{k\in\mathsf{Z}} \sum_{j\in\mathbb{Z}} |\alpha_j \beta_{k-j}| \mathbb{E}\,|X_{t-k}| \\ &\leq \sup_{s\in\mathbb{Z}} \mathbb{E}\,|X_s| \times \sum_{s\in\mathsf{Z}} |\alpha_s| \times \sum_{s\in\mathbb{Z}} |\beta_s| . \end{aligned}$$

Hence we obtain

$$\sum_{k\in\mathsf{Z}} \sum_{j\in\mathbb{Z}} |\alpha_j \beta_{k-j} X_{t-k}| < \infty \quad \text{a.s.} .$$

We can thus apply Fubini's Theorem and get that

$$\begin{aligned} W_t &= \sum_{j\in\mathbb{Z}} \alpha_j \left(\sum_{k\in\mathsf{Z}} \beta_{k-j} X_{t-k} \right) \quad \text{a.s.} \\ &= \sum_{j\in\mathbb{Z}} \alpha_j Y_{t-j} \quad \text{a.s.} \\ &= Z_t \quad \text{a.s.} \end{aligned}$$

Hence the result. □

An immediate consequence of Lemma 3.2.1 is that F_α and F_β commute, since the convolution product $\star$ commute in ℓ^1. Another important consequence is that inverting a linear filter F_α by another linear filter F_β, that is, finding $\beta \in \ell^1$ such that $\mathrm{F}_\alpha \circ \mathrm{F}_\beta$ is the identity operator, is equivalent to finding $\beta \in \ell^1$ such that $\alpha \star \beta = \mathrm{e}_0$, where e_0 is the impulsion sequence defined by

$$\mathrm{e}_{0,t} = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Now define the Fourier series α^* and β^* as in (3.9). It is easy to show that, for all $\alpha, \beta \in \ell^1$,

$$(\alpha \star \beta)^* = \alpha^* \times \beta^* \, .$$

Consequently, we have

$$\alpha \star \beta = \mathrm{e}_0 \Leftrightarrow \alpha^* \times \beta^* = 1 \, . \tag{3.10}$$

Let us sum up these findings in the following proposition.

Proposition 3.2.2. *Let $\alpha, \beta \in \ell^1$. Define the Fourier series α^* and β^* as in (3.9) and suppose that $\alpha^* \times \beta^* = 1$. Then, for all random process $X = (X_t)_{t \in \mathbb{Z}}$ satisfying (3.4), we have*

$$\mathrm{F}_\alpha \circ \mathrm{F}_\beta(X) = \mathrm{F}_\beta \circ \mathrm{F}_\alpha(X) = X \quad a.s.$$

Of course, not all $\alpha \in \ell^1$ defines a filter F_α which is "invertible" in the sense of Proposition 3.2.2, that is admits a $\beta \in \ell^1$ such that $\alpha \star \beta = \mathrm{e}_0$. Nevertheless, the case where F_α is a FIR filter can be completely described by the following lemma.

Lemma 3.2.3. *Let P and Q be two polynomials with complex coefficients with no common roots. Assume that $Q(0) = 1$ and that Q does not vanish on the unit circle*

$$\mathbb{U} = \{z \in \mathbb{C} \, : \, |z| = 1\} \, .$$

The rational function P/Q admits the following uniformly convergent series expansion

$$\frac{P}{Q}(z) = \sum_{t \in \mathsf{Z}} \psi_t z^t \, , \tag{3.11}$$

on the ring

$$R_{\delta_1, \delta_2} = \{z \in \mathbb{C} \, , \delta_1 < |z| < \delta_2\} \, ,$$

where $\psi \in \ell^1$ and

$$\begin{aligned} \delta_1 &= \max\{|z| \ : \ z \in \mathbb{C}, |z| < 1, \, Q(z) = 0\} \\ \delta_2 &= \min\{|z| \ : \ z \in \mathbb{C}, |z| > 1, \, Q(z) = 0\} \, . \end{aligned}$$

with the convention $\max(\emptyset) = 0$ and $\min(\emptyset) = \infty$.

If P and Q have real valued coefficient, so has ψ.

Moreover, the two following assertions hold and provides the asymptotic behavior of ψ_t as $t \to \pm\infty$.

(i) We have $\psi_t = 0$ for all $t < 0$ if and only if $\delta_1 = 0$, that is, if and only if Q does not vanish on the unit disk $\Delta_1 = \{z \in \mathbb{C} \ : \ |z| \leq 1\}$. If it is not the case, then, for any $\eta \in (0, \delta_1)$, $\psi_t = O(\eta^{-t})$ as $t \to -\infty$.

(ii) We have $\psi_t = 0$ for all $t > \deg(P) - \deg(Q)$ if and only if $\delta_2 = \infty$, that is, if and only if Q does not vanish out of the unit disk Δ_1. If it is not the case, then, for any $\eta \in (0, 1/\delta_2)$, $\psi_t = O(\eta^t)$ as $t \to \infty$.

Proof. By the partial fraction decomposition of the P/Q, one can first solve the case where Q has degree 1. The details of the proof is left to the reader (see Exercise 3.6). □

The series expansion (3.11) extends the classical expansion of power series to a two-sided sum. It is called a *Laurent series expansion.*

Applying Lemma 3.2.3 to solve Proposition 3.2.2 in the special case (3.3), we get the following result.

Corollary 3.2.4. *Under the assumptions of Lemma 3.2.3, we have, for all random process $X = (X_t)_{t\in\mathbb{Z}}$ satisfying (3.4),*

$$\mathrm{F}_\psi \circ [Q(\mathrm{B})](X) = [P(\mathrm{B})](X) \ ,$$

where ψ is the unique sequence in ℓ^1 that satisfies (3.11) for all $z \in \mathbb{U}$ (the unit circle).

Proof. The only fact to show is that (3.11) on $z \in \mathbb{U}$ uniquely defines ψ. (We already know that ψ exists and belongs to ℓ^1 from Lemma 3.2.3). This fact follows from the inverse Fourier transform. Namely, for all $\psi \in \ell^1$, defining ψ^* as in (3.9), it is easy to show that, for all $t \in \mathbb{Z}$,

$$\psi_t = \frac{1}{2\pi} \int_{\mathbb{T}} \psi^*(\lambda)\, \mathrm{e}^{\mathrm{i}t\lambda}\, \mathrm{d}\lambda \ .$$

Hence the result. □

Applying Corollary 3.2.4 with $P = 1$ allows us to derive the inverse filter of any FIR filter of the form $Q(\mathrm{B})$.

Another interesting application of Corollary 3.2.4 is to derive nontrivial filters whose effects on the spectral density is a multiplication by a constant; they are called *all-pass filters.*

Definition 3.2.2 (All-pass filters)**.** *Let $\psi \in \ell^1$. The linear filter F_ψ is called an all-pass filter if there exists $c > 0$ such that, for all z on the unit circle Γ_1,*

$$\left| \sum_{k\in\mathbb{Z}} \psi_k z^k \right| = c \ .$$

An interesting obvious property of these filters is the following.

Lemma 3.2.5. *Let $\psi \in \ell^1$ such that F_ψ is an all-pass filter. Then if Z is a weak white noise, so is $\mathrm{F}_\psi(Z)$.*

Other type of filters satisfy this property, such as the time reversion operator, see Example 2.1.6.

Example 3.2.1 (All-pass filter, a trivial case)**.** *Any filter of the form aB^k with $a \in \mathbb{C}$ and $k \in \mathbb{Z}$ is an all-pass filter, since it corresponds to F_ψ with $\psi_l = 0$ for all $l \neq k$ and $\psi_k = a$.*

A more interesting example is obtained starting from a given polynomial Q.

Example 3.2.2 (All-pass filter inverting the roots moduli)**.** *Let Q be a polynomial such that $Q(0) = 1$, so that*

$$Q(z) = \prod_{k=1}^{p} (1 - \nu_k z) \, ,$$

where p is the degree of Q and $\nu_1, \ldots, \nu_p$ are the reciprocals of its roots. Define the polynomial

$$\tilde{Q}(z) = \prod_{k=1}^{p} (1 - \overline{\nu_k^{-1}} z) \, .$$

Assume that Q does not vanish on the unit circle Γ_1, so that the same holds for $\tilde{Q}$. Then we have, for all z on Γ_1,

$$\left| \frac{Q(z)}{\tilde{Q}(z)} \right|^2 = \prod_{k=1}^{p} |\nu_k|^2 \, . \tag{3.12}$$

By Corollary 3.2.4, there exists a unique $\tilde{\psi} \in \ell^1$ such that

$$\frac{1}{\tilde{Q}}(z) = \sum_{t \in \mathsf{Z}} \tilde{\psi}_t z^t \, , \tag{3.13}$$

and we have $\mathrm{F}_{\tilde{\psi}} \circ [\tilde{Q}(\mathrm{B})](X) = X$ for all $X = (X_t)_{t \in \mathbb{Z}}$ satisfying (3.4). Define $\phi \in \ell^1$ such that

$$\mathrm{F}_{\phi} = \mathrm{F}_{\tilde{\psi}} \circ [Q(\mathrm{B})] \, .$$

As a consequence of (3.12) and (3.13), the filter F_{ϕ} is an all-pass filter and satisfies

$$\mathrm{F}_{\phi} \circ [\tilde{Q}(B)] = [Q(B)] \, . \tag{3.14}$$

Proceeding similarly with $\tilde{Q}$ replacing Q (and Q replacing $\tilde{Q}$), we obtain $\tilde{\phi} \in \ell^1$ such that $\mathrm{F}_{\tilde{\phi}}$ is an all-pass filter and satisfies

$$\mathrm{F}_{\tilde{\phi}} \circ [Q(B)] = [\tilde{Q}(B)] \, . \tag{3.15}$$

Moreover, we have $\phi \star \tilde{\phi} = \mathrm{e}_0$, so that

$$\mathrm{F}_{\phi} \circ \mathrm{F}_{\tilde{\phi}} = \mathbf{1} \, . \tag{3.16}$$

Here $\mathbf{1}$ denotes the identity operator and all operators above are defined on the class of all processes that satisfy (3.4) (in particular on the class of weakly stationary processes). Observe moreover that if Q is a polynomial with real coefficients, then so is $\tilde{Q}$ and ϕ also takes its values in $\mathbb{R}$.

3.3 Definition of ARMA processes

In the following we take the convention that ARMA processes are centered. To define a *noncentered* ARMA process, just add a constant to a centered ARMA process. We will work with complex valued ARMA processes for convenience, although in practice, for modelling purposes, one usually works with real valued ARMA processes. From a theoretical point of view, there is not much difference between the two settings, except concerning existence results: it can be a bit harder to prove the existence of a real-valued process than a complex-valued process.

3.3.1 MA(q) processes

Definition 3.3.1 (MA(q) processes)**.** *A random process $X = (X_t)_{t\in\mathbb{Z}}$ is called a* moving average *process of order q (MA(q)) with coefficients $\theta_1, \dots, \theta_q$ if it satisfies the MA(q) equation*

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \ , \tag{3.17}$$

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

In other word $X = \mathrm{F}_\alpha(Z)$, where F_α is a FIR filter with coefficients

$$\alpha_t = \begin{cases} 1 & \text{if } t = 0, \\ \theta_k & \text{if } t = 1, \dots, q, \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

Equivalently, we can write

$$X = [\Theta(\mathrm{B})](Z) \ ,$$

where B is the Backshift operator and Θ is the polynomial defined by $\Theta(z) = 1 + \sum_{k=1}^{p} \theta_k z^k$.

Hence it is a linear process with short memory, and by Corollary 3.1.3, it is a centered weakly stationary process with autocovariance function given by

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{t=0}^{q-h} \theta_k \overline{\theta}_{k+h}, & \text{if } 0 \le h \le q \ , \\ \sigma^2 \sum_{t=0}^{q+h} \overline{\theta}_k \theta_{k-h}, & \text{if } -q \le h \le 0 \ , \\ 0, & \text{otherwise} \ , \end{cases} \tag{3.19}$$

and with spectral density function given by

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| 1 + \sum_{k=1}^{q} \theta_k \mathrm{e}^{-\mathrm{i}k\lambda} \right|^2 \ .$$

We already mentioned the MA(1) process in Example 2.3.1, and displayed its spectral density in Figure 2.3.

3.3.2 AR(p) processes

Definition 3.3.2 (AR(p) processes)**.** *A random process $X = (X_t)_{t\in\mathbb{Z}}$ is called an* autoregressive *process of order p (AR(p)) with coefficients $\phi_1, \dots, \phi_p$ if it satisfies the AR(p) equation*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \ , \tag{3.20}$$

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

Observe that (3.20) looks like a regression model where the regressors are given by the p past values of the process. Hence the term "autoregressive". This is also the reason why the AR processes are so popular for modelling purposes.

In contrast with MA process, this sole definition does not guaranty that X is weakly stationary. In fact, as soon as $\phi_k \neq 0$ for some k (otherwise $X = Z$), this equation has clearly an infinite set of solutions! It suffices to choose an arbitrary set of initial conditions

$(X_0, X_{-1}, \ldots, X_{1-p}$ (possible independently of the process Z) and to compute X_t by iterating (3.20) for $t \geq 1$ and by iterating the backward equation

$$X_{t-p} = \frac{1}{\phi_p} X_t - \frac{\phi_1}{\phi_p} X_{t-1} - \cdots - \frac{\phi_{p-1}}{\phi_p} X_{t-p+1} + Z_t \ , \tag{3.21}$$

for $t \leq -1$.

Nevertheless, for well chosen AR coefficients $\phi_1, \ldots, \phi_p$, there a unique weakly stationary process that satisfies the AR(p) equation (3.20). Unless otherwise stated, *the* AR(p) process defined by an AR(p) equation will always be taken as this weakly stationary solution.

To better understand this point of view, let us consider the case $p = 1$,

$$X_t = \phi X_{t-1} + Z_t \ . \tag{3.22}$$

By iterating this equation, we get

$$X_t = \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j} \ . \tag{3.23}$$

Let us first assume that $|\phi| < 1$. If we assume X to be weakly stationary then, taking the limit (in the L^2 sense) as $k \to \infty$, we get

$$X = \mathrm{F}_\psi(Z) \ ,$$

where

$$\psi_t = \begin{cases} \phi^t & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is simple verification to check that this weakly stationary process is indeed a solution to the AR(1) equation (3.22). So we have shown our claim when $|\phi| < 1$.

If $|\phi| > 1$, it is easy to adapt the previous proof by using the backward recursion (3.21) in the case $p = 1$. In this case, we obtain again that there is unique weakly stationary solution to the AR(1) equation, and it is given by $X = \mathrm{F}_\psi(Z)$, this time with

$$\psi_t = \begin{cases} \phi^t & \text{if } t \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, if $|\phi| = 1$, rewriting (3.23) as

$$X_t - \phi^k X_{t-k} = \sum_{j=0}^{k-1} \phi^j Z_{t-j} \ ,$$

we observe that the right-hand side has variance $k\sigma^2$, while the left-hand side has variance at most $2(\mathrm{Var}\,(X_t) + \mathrm{Var}\,(X_{t-k}))$ hence would be bounded if X were weakly stationary. We conclude that in this case, there is no weakly stationary solution to the AR(1) equation.

In conclusion we have shown the following result in the case $p = 1$.

Theorem 3.3.1 (Existence and uniqueness of a weakly stationary solution of the AR(p) equation)**.** *Let $Z \sim \mathrm{WN}(0, \sigma^2)$ with $\sigma^2 > 0$ and $\phi_1, \dots, \phi_p \in \mathbb{C}$. Define the polynomial*

$$\Phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k \, .$$

Then the AR(p) equation (3.20) has a unique weakly stationary solution X if and only if Φ does not vanish on the unit circle $\mathbb{U}$. Moreover, in this case, we have $X = \mathrm{F}_\psi(Z)$, where $\psi \in \ell^1$ is uniquely defined by

$$\sum_{t \in \mathbb{Z}} \psi_t z^t = \frac{1}{\Phi(z)} \quad \text{on} \quad z \in \mathbb{U} \, .$$

The proof in the general case is omitted since we will treat below the more general ARMA recurrence equations, see Theorem 3.3.2.

Let us just mention that it easily follows from our result on the inversion of FIR filters (see Corollary 3.2.4) by observing that, as for MA processes, the AR(p) equation can be interpreted as a FIR filter equation, namely, $Z = \mathrm{F}_\beta(X)$, where F_β is a FIR filter with coefficients

$$\beta_t = \begin{cases} 1 & \text{if } t = 0, \\ -\phi_t & \text{if } t = 1, \dots, p, \\ 0 & \text{otherwise.} \end{cases} \tag{3.24}$$

Or, equivalently, $Z = [\Phi(\mathrm{B})](X)$.

3.3.3 ARMA(p, q) processes

ARMA(p, q) processes is an extension both of AR(p) and MA(q) processes.

Definition 3.3.3 (ARMA(p, q) processes)**.** *A random process $X = (X_t)_{t \in \mathbb{Z}}$ is called an* autoregressive moving average *process of order (p, q) (ARMA(p, q)) with AR coefficients $\phi_1, \dots, \phi_p$ and MA coefficients $\theta_1, \dots, \theta_q$ if it satisfies the ARMA(p, q) equation*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \, , \tag{3.25}$$

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

As discussed for the AR(p) equation, again the ARMA(p, q) equation has an infinite set of solutions, but at most one that is weakly stationary and this happens for well chosen AR coefficients.

Before stating this result, let us recall how the ARMA equation can be rewritten using linear filer operators. The ARMA(p, q) equation can be written as

$$\Phi(\mathrm{B})(X) = \Theta(\mathrm{B})(Z) \, , \tag{3.26}$$

where B is the Backshift operator and Φ and Θ are the polynomials defined by

$$\Phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k \quad \text{and} \quad \Theta(z) = 1 + \sum_{k=1}^{p} \theta_k z^k \, . \tag{3.27}$$

To avoid treating useless particular cases, it is natural to assume that Φ and Θ have no common roots. Otherwise, factorizing these polynomials, we see that the same operators apply to both sides of (3.26).

Theorem 3.3.2 (Existence and uniqueness of a weakly stationary solution of the ARMA(p, q) equation)**.** *Let $Z \sim \mathrm{WN}(0, \sigma^2)$ with $\sigma^2 > 0$ and $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q \in \mathbb{C}$. Assume that the polynomials Φ and Θ defined by (3.27) have no common roots. Then the ARMA(p, q) equation (3.20) has a unique weakly stationary solution X if and only if Φ does not vanish on the unit circle $\mathbb{U}$. Moreover, in this case, we have $X = \mathrm{F}_\psi(Z)$, where $\psi \in \ell^1$ is uniquely defined by*

$$\sum_{t\in\mathbb{Z}} \psi_t z^t = \frac{\Theta}{\Phi}(z) \quad \text{on} \quad z \in \mathbb{U}\,. \tag{3.28}$$

As a consequence, X admits a spectral density function given by

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{\Theta}{\Phi}(\mathrm{e}^{-\mathrm{i}\lambda})\right|^2\,. \tag{3.29}$$

Remark 3.3.1. *In fact (3.28) holds in the ring $\{z \in \mathbb{C}\,, \delta_1 < |z| < \delta_2\}$, where $\delta_1 = \max\{z \in \mathbb{C}, |z| < 1, \phi(z) = 0\}$ and $\delta_2 = \min\{z \in \mathbb{C}, |z| > 1, \phi(z) = 0\}$.*

Proof of Theorem 3.3.2. We first suppose that Φ does not vanish on the unit circle. Since the ARMA(p, q) equation can be rewritten as

$$[\Phi(\mathrm{B})](X) = [\Theta(\mathrm{B})](Z)\,,$$

existence and uniqueness of a weakly stationary solution directly follows from Corollary 3.2.4: setting $X = \mathrm{F}_\psi(Z)$ gives the existence; applying F_ξ to both sides of this equation gives the uniqueness, where $\xi \in \ell^1$ satisfies

$$\sum_{t\in\mathbb{Z}} \xi_t z^t = \frac{1}{\Phi(z)} \quad \text{on} \quad z \in \mathbb{U}\,.$$

(we apply Corollary 3.2.4 with $P = 1$). The spectral density function expression (3.29) then follows from Theorem 3.1.2.

It only remains to show that if Φ does vanish on the unit circle, then the ARMA(p, q) equation does not admits a weakly stationary solution. Let $\lambda_0 \in \mathbb{T}$ such that $\mathrm{e}^{-\mathrm{i}\lambda_0}$ is a root of Φ and let X be a weakly stationary process with spectral measure ν. Using Theorem 3.1.2, it follows that $[\Phi(\mathrm{B})](X)$ has a spectral measure ν' such that, for all $\epsilon > 0$,

$$\begin{aligned}\nu'([\lambda_0 - \epsilon, \lambda_0 + \epsilon]) &= \int_{[\lambda_0-\epsilon,\lambda_0+\epsilon]} |\Phi(\mathrm{e}^{-\mathrm{i}\lambda})|^2\, \nu(\mathrm{d}\lambda) \\ &\le C\,\epsilon^2\, \nu([\lambda_0 - \epsilon, \lambda_0 + \epsilon]) \\ &= O(\epsilon^2)\,.\end{aligned}$$

On the other hand, $[\Theta(\mathrm{B})](Z)$ has a continuous spectral density which does not vanish at λ_0, since Θ has no common roots with Φ and thus does not vanish at $\mathrm{e}^{-\mathrm{i}\lambda_0}$, so its spectral measure applied to the same set $[\lambda_0 - \epsilon, \lambda_0 + \epsilon]$ is lower bounded by $c\epsilon$ with $c > 0$. Hence we cannot have $[\Phi(\mathrm{B})](X) = [\Theta(\mathrm{B})](Z)$, which concludes the proof. □

3.4 Representations of an ARMA(p, q) process

In view of Definition 3.1.1 and Definition 3.2.1,

Definition 3.4.1 (Representations of ARMA(p,q) processes)**.** *If the ARMA equation (3.25) has a weakly stationary solution $X = \mathrm{F}_\psi(Z)$, it is said to provide*

(i) *a* causal representation *of X if F_ψ is a causal filter,*

(ii) *an* invertible representation *of X if $\mathrm{F}_\psi(Z)$ is an invertible representation and its inverse filter is causal,*

(iii) *a* canonical representation *of X if $\mathrm{F}_\psi(Z)$ is a causal and invertible representation.*

We have the following result.

Theorem 3.4.1. *Under the assumptions and notation of Theorem 3.3.2, the ARMA equation (3.25) provides*

(i) *a* causal representation *of X if and only if Φ does not vanish on the unit closed disk Δ_1,*

(ii) *an* invertible representation *of X if and only if Θ does not vanish on the unit closed disk Δ_1,*

(iii) *a* canonical representation *of X if and only if neither Φ nor Θ does vanish on the unit closed disk Δ_1.*

Proof. The characterization of the causality of F_ψ directly follows from the definition of ψ in Theorem 3.3.2 and from Lemma 3.2.3.

The second equivalence is obtained similarly by inverting the roles of Φ and Θ.

The third equivalence follows from the first two. □

We shall see in the following that a canonical representation is very useful to derive the innovation process of an ARMA process X. Applying the all-pass filters derived in Example 3.2.2, we easily get the following result.

Theorem 3.4.2. *Let X be the weakly stationary solution of the ARMA equation (3.25), where Φ and Θ defined by (3.27) have no common roots and no roots on the unit circles. Then there exists AR coefficients $\tilde{\phi}_1, \ldots, \tilde{\phi}_p$ and MA coefficients $\tilde{\theta}_1, \ldots, \tilde{\theta}_q$ and $\tilde{Z} \sim \mathrm{WN}(0, \sigma^2)$ such that X satisfies the ARMA(p,q) equation*

$$X_t = \tilde{\phi}_1 X_{t-1} + \cdots + \tilde{\phi}_p X_{t-p} + \tilde{Z}_t + \tilde{\theta}_1 \tilde{Z}_{t-1} + \cdots + \tilde{\theta}_q \tilde{Z}_{t-q} \,, \tag{3.30}$$

and the corresponding polynomials $\tilde{\Phi}$ and $\tilde{\Theta}$ do not vanish on the unit closed disk Δ_1. In particular, (3.30) is a canonical representation of X. Moreover, if the original AR and MA coefficients ϕ_k's and θ_k's are real, so are the canonical ones $\tilde{\phi}_k$'s and $\tilde{\theta}_k$'s.

Proof. We may write $\Phi = P \times Q$, where P has its roots out of Δ_1 and Q in the interior of Δ_1 and $P(0) = Q(0) = 1$. Proceeding as in Example 3.2.2, we obtain $\phi, \tilde{\phi} \in \ell^1$ such that (3.14) holds and F_ϕ is an all-pass filter. Applying F_ϕ to both sides of (3.26) and using (3.14), we get

$$\tilde{\Phi}(B)(X) = \Theta(B) \circ \mathrm{F}_\phi(Z) \,,$$

where $\tilde{\Phi} = P \times \tilde{Q}$ is a polynomial with same degree as Φ and all its roots out of Δ_1. We can proceed similarly with the polynomial Θ and obtain a polynomial $\tilde{\Theta}$ with same degree as Θ and roots out of Δ_1 and $\tilde{\phi} \in \ell^1$ such that F_ϕ is an all-pass filter and

$$\Theta(B) = \tilde{\Theta}(B) \circ \mathrm{F}_{\tilde{\phi}} \ .$$

As a consequence we obtain that X is solution to the equation

$$\tilde{\Phi}(B)(X) = \tilde{\Theta}(B) \circ \mathrm{F}_{\tilde{\phi}} \circ \mathrm{F}_\phi(Z) \ .$$

Now, by Lemma 3.2.5, we know that $\mathrm{F}_{\tilde{\phi}} \circ \mathrm{F}_\phi(Z)$ is a white noise. Hence the previous displayed equation is an ARMA equation that admits a unique weakly stationary solution, which is X. Moreover, by construction, it provides a canonical representation of X. □

Theorem 3.4.2 is a very important result as it provides a canonical representation of any ARMA process X, provided that the polynomials of the original ARMA equation do not vanish on the unit circle.

3.5 Innovations of ARMA processes

Interestingly, a canonical representation of an ARMA process provides the innovations of the process, as shown by the following result.

Theorem 3.5.1. *Let X be the weakly stationary solution to a canonical ARMA(p, q) equation of the form*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \ ,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$. Then Z is the innovation process of X.

Proof. By definition of the canonical representation, there exists $\psi, \tilde{\psi} \in \ell^1$ such that $\psi_k = \tilde{\psi}_k = 0$ for all $k < 0$, $X = \mathrm{F}_\psi(Z)$ and $Z = \mathrm{F}_{\tilde{\psi}}(X)$. We deduce that, for all $t \in \mathbb{Z}$, $\mathcal{H}_t^Z = \mathcal{H}_t^X$. Consequently, for all $t \in \mathbb{Z}$,

$$\hat{X}_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \in \mathcal{H}_{t-1}^X \ ,$$

and

$$X_t - \hat{X}_t = Z_t \in \mathcal{H}_t^Z \perp \mathcal{H}_{t-1}^Z = \mathcal{H}_{t-1}^X \ .$$

Hence, by Theorem A.4.1, we obtain that

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1}^X\right) = \hat{X}_t \ .$$

Hence the result. □

From (3.19), we see that an MA(q) process has an Autocovariance function $\gamma(h)$ which vanishes for all $|h| > q$. A very important result is the converse implication. Its proof relies on the construction of the innovation process from the assumption on the autocovariance function γ.

Theorem 3.5.2. *Let X be a centered weakly stationary process with autocovariance function γ. Then X is an MA(q) process if and only if $\gamma(h) = 0$ for all $|h| > q$.*

Proof. The "only if" part is already known. We thus show the "if" part, that is, we take a centered weakly stationary process X with autocovariance function γ, assume that $\gamma(h) = 0$ for all $|h| > q$, and show that it is an MA(q) process.

Let $(\epsilon_t)_{t\in\mathbb{Z}}$ be the innovation process of X, thus it is a white noise WN$(0, \sigma^2)$, see Section 2.6. Since $\gamma(h) = 0$ for all $|h| > q$, we have $X_t \perp \mathcal{H}^X_{t-q-1}$ for all t. Observing that

$$\mathcal{H}^X_{t-1} = \mathcal{H}^X_{t-q-1} \overset{\perp}{\oplus} \mathrm{Span}\,(\epsilon_{t-q}, \ldots, \epsilon_{t-1}) \ ,$$

by Property (*vii*) in Proposition A.4.2, we obtain that

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}^X_{t-1}\right) = \mathrm{proj}\left(X_t \middle| \mathrm{Span}\,(\epsilon_{t-q}, \ldots, \epsilon_{t-1})\right) \ ,$$

and thus, either $\sigma^2 = 0$ and $X_t = 0$ a.s. (a very trivial MA process) or X is regular and we have

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}^X_{t-1}\right) = \sum_{k=1}^{q} \frac{\langle X_t, \epsilon_{t-k}\rangle}{\sigma^2}\, \epsilon_{t-k} \ .$$

where the coefficient in front of each ϵ_{t-k} does not depend on t, but only on k (see (2.22). Let us presently denote it by θ_k. Since $X_t = \mathrm{proj}\left(X_t \middle| \mathcal{H}^X_{t-1}\right) + \epsilon_t$, we finally get that X is solution of (3.17) with the white noise Z replaced by the innovation process ϵ (which also is a white noise). Hence X is an MA(q) process. □

Remark 3.5.1. *We have authorized ARMA processes to be complex valued. The question arises whether the "if part" of Theorem 3.5.2 continues to hold for real MA processes. Inspecting the proof of this result, the answer is yes. If one start with a real valued process X, then the prediction coefficients and the innovation process are real valued, and so are the coefficients $\theta_1, \ldots, \theta_q$ defined in this proof.*

To conclude with the innovations of ARMA processes, we show the following result, which is a specialization of Theorem 3.5.1 to the case of AR processes.

Theorem 3.5.3. *Let X be a weakly stationary AR(p) process with causal representation*

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

where $Z \sim \mathrm{WN}(0, \sigma^2)$. Then, for all $m \geq p$, the prediction coefficients are given by

$$\phi^+_p = [\phi_1, \ldots, \phi_p, \underbrace{0, \ldots, 0}_{m-p}]^T \ ,$$

that is, for all $t \in \mathbb{Z}$,

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,m}\right) = \sum_{k=1}^{p} \phi_k X_{t-k} \ .$$

In particular the prediction error of order m is Z_t and has variance σ^2 and thus is constant for all $m \geq p$.

Proof. The proof follows that of Theorem 3.5.1. □

This property provides a characterization of AR(p) processes as simple as that provided for MA(q) processes in Theorem 3.5.2. It relies on the following definition.

Definition 3.5.1 (Partial autocorrelation function)**.** *Let X be a weakly stationary process. The* partial autocorrelation function *of X is the function defined by*

$$\kappa(p) = \phi^+_{p,p}, \quad p = 1, 2, \dots$$

where $\phi^+_p = \left(\phi^+_{k,p}\right)_{k=1,\dots,p}$ denote the prediction coefficients of X, that is, for all $t \in \mathbb{Z}$,

$$\operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,p}\right) = \sum_{k=1}^{p} \phi^+_{k,p} X_{t-k} \ ,$$

with the convention that $\kappa(p) = 0$ if this equation does not defines uniquely ϕ^+_p, that is, if Γ^+_p is not invertible.

We see from Theorem 3.5.3 that if X is an AR process, then its partial autocorrelation function vanishes for all $m > p$. It is in fact a characterization of AR processes, as shown by the following result.

Theorem 3.5.4. *Let X be a centered weakly stationary process with partial autocorrelation function κ. Then X is an AR(p) process if and only if $\kappa(m) = 0$ for all $m > p$.*

Proof. The "only if" part is a consequence of Theorem 3.5.3.

Let us show the "if" part. Let X be a centered weakly stationary process with partial autocorrelation function κ such that $\kappa(m) = 0$ for all $m > p$. This implies that, for all such m and all $t \in \mathbb{Z}$,

$$\operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,m}\right) \in \mathcal{H}^X_{t-1,m-1} \ ,$$

which implies that

$$\operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,m}\right) = \operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,m-1}\right) \ ,$$

and, iterating in m,

$$\operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,m}\right) = \operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,p}\right) \ .$$

Letting $m \to \infty$, by (2.16), we get that

$$\operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1}\right) = \operatorname{proj}\left(X_t | \mathcal{H}^X_{t-1,p}\right) = \sum_{k=1}^{p} \phi_p X_{t-k} \ ,$$

where $\phi_1, \dots, \phi_p$ are the prediction coefficients of order p. Denote by Z the innovation process of X, then Z is a white noise (see Corollary 2.6.1) and X satisfies the AR(p) equation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \ .$$

Hence the result. □

3.6 Autocovariance function of ARMA processes

The spectral density of an ARMA process is easily obtained from the AR and MA coefficients by (3.29).

We now explain in this section how to compute the autocovariance function of an ARMA process. For this purpose we assume in this section that X is the weakly stationary solution of a causal ARMA(p,q) equation of the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \ , \tag{3.31}$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$. Note that, whenever a stationary solution exists, a causal representation of the ARMA equation can be found, see the first part of the proof of Theorem 3.4.2.

Algorithm 3: Computation of the filter coefficients and the autocovariance function from a causal ARMA representation.

Data: AR and MA coefficients $\phi_1, \ldots, \phi_r$, $\theta_1, \ldots, \theta_r$, and variance σ^2 of the white noise.

Result: Causal filter coefficients $(\psi_k)_{k\geq 0}$ and autocovariance function γ.

Step 1 Initialization: set $\psi_0 = 1$.

for $k = 1, 2, \ldots, r$ **do**

Compute

$$\psi_k = \theta_k + \sum_{j=1}^{k} \psi_{k-j}\phi_j \ . \tag{3.32}$$

end

for $k = r+1, r+2, \ldots$ **do**

Compute

$$\psi_k = \sum_{j=1}^{r} \psi_{k-j}\phi_j \ . \tag{3.33}$$

end

Step 2 **for** $\tau = 0, 1, 2, \ldots$ **do**

Compute

$$\gamma(\tau) = \sigma^2 \sum_{k=0}^{\infty} \overline{\psi_k}\psi_{k+\tau} \ . \tag{3.34}$$

end

and **for** $\tau = -1, -2, \ldots$ **do**

Set

$$\gamma(\tau) = \overline{\gamma(-\tau)} \ .$$

end

Theorem 3.6.1. *Let X be the weakly stationary solution of the ARMA(p,q) equation (3.31), which is assumed to be a causal representation, that is, for all $z \in \mathbb{C}$ such that $|z| \leq 1$,*

$$1 - \sum_{k=1}^{p} \phi_k z^k \neq 0 \ .$$

Define $r = \max(p,q)$ and set $\theta_j = 0$ for $q < j \leq r$ or $\phi_j = 0$ for $p < j \leq r$. Then Algorithm 3 applies.

Proof. Because the representation is causal, we know that the solution $\psi \in \ell^1$ of the equation (3.28) satisfies $\psi_k = 0$ for all $k < 0$. Moreover, by Lemma 3.2.3 and (3.10), this equation can be interpreted as the convolution equation

$$\psi \star \phi = \theta \ ,$$

where ϕ and θ here denote the sequences associated to the polynomial Φ and Θ by the relations

$$\phi^*(\lambda) = \Phi(\mathrm{e}^{-\mathrm{i}\lambda}) \ ,$$

and

$$\theta^*(\lambda) = \Theta(\mathrm{e}^{-\mathrm{i}\lambda}) \ ,$$

Because ψ is one-sided and ϕ has a finite support, and using the definition of r, we easily get

$$\begin{aligned}
&\psi_0 = 1 \\
&\psi_1 = \theta_1 + \psi_0\phi_1 \\
&\psi_2 = \theta_2 + \psi_0\phi_2 + \psi_1\phi_1 \\
&\vdots \\
&\psi_r = \theta_r + \sum_{j=1}^{r} \psi_{k-j}\phi_j \\
&\psi_{r+1} = \sum_{j=1}^{r} \psi_{r+1-j}\phi_j \\
&\vdots
\end{aligned}$$

that is, (3.32) and (3.33) hold, which achieves the proof of **Step 1**.

The computations of **Step 2** directly follow from Corollary 3.1.3 in the case where ψ vanishes on $\mathbb{Z}_-$, which concludes the proof. □

Observe that Algorithm 3 has to be performed formally in the sense that it involves infinite recursions and sums, even if only a finite number of values of the autocovariance function is computed. In contrast the next algorithm can be performed numerically : only a finite number of operations is necessary for computing a finite number of covariance coefficients.

Algorithm 4: Computation of the autocovariance function from a causal ARMA representation.

Data: AR and MA coefficients $\phi_1, \dots, \phi_r$, $\theta_1, \dots, \theta_r$, and variance σ^2 of the white noise, a lag m.

Result: Causal filter coefficients ψ_k for $k = 0, \dots, r$ and autocovariance function $\gamma(\tau)$ for $\tau = -m, \dots, m$.

Step 1 Initialization: set $\psi_0 = 1$.

for $k = 1, 2, \dots, r$ **do**

Compute ψ_k by applying (3.32).

end

Step 2 Using that $\gamma(-j) = \overline{\gamma(j)}$ for all j and setting $\theta_0 = 1$, solve the linear system

$$\gamma(\tau) - \phi_1\gamma(\tau-1) - \cdots - \phi_r\gamma(\tau-r) = \sigma^2 \sum_{\tau \le j \le r} \theta_j \overline{\psi}_{j-\tau}, \quad 0 \le \tau \le r\,, \tag{3.35}$$

in $\gamma(\tau)$, $\tau = 0, 1, 2, \dots, r$.

Step 3 Then apply the following induction.

for $\tau = r+1, r+2, \dots, m$ **do**

Compute

$$\gamma(\tau) = \phi_1\gamma(\tau-1) + \cdots + \phi_r\gamma(\tau-r)\,. \tag{3.36}$$

end

for $\tau = -1, -2, \dots, -m$ **do**

Set

$$\gamma(\tau) = \overline{\gamma(-\tau)}\,.$$

end

Theorem 3.6.2. *Under the same assumptions as Theorem 3.6.1, Algorithm 3 applies.*

Proof. The proof of **Step 1** is already given in the proof of Theorem 3.6.1. Observe that, by causality, we have $X_t = \sum_{\ell \ge 0} \psi_\ell Z_{t-\ell}$ and thus, for all $t, \tau \in \mathbb{Z}$ and $j = 0, \dots, r$,

$$\operatorname{Cov}(Z_{t-j}, X_{t-\tau}) = \begin{cases} \sigma^2 \overline{\psi}_{j-\tau} & \text{if } j \ge \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Now by (3.31), taking the covariance both sides with $X_{t-\tau}$, we get (3.35) for $0 \le \tau \le r$ and (3.36) for $\tau \ge r+1$. □

3.7 Exercises

Exercise 3.1. Define $Y = (Y_t)_{t\in\mathbb{Z}}$ by

$$Y_t = \beta t + S_t + X_t, \quad t \in \mathbb{Z}\,,$$

where $\beta \in \mathbb{R}$, $S = (S_t)_{t\in\mathbb{Z}}$ is a 4-periodic weakly stationary process and $X = (X_t)_{t\in\mathbb{Z}}$ is a weakly stationary process such that X and S are uncorrelated.

1. Is Y weakly stationary ?
2. Which property is satisfied by the covariance function of S ? Define $\left(\bar{S}_t\right)_{t\in\mathbb{Z}}$ as the process obtained by applying the operator $\mathbf{1} + \mathrm{B} + \mathrm{B}^2 + \mathrm{B}^3$ to S, where B denotes the shift operator. What can be said about $\bar{S}$?
3. Consider now Z obtained by applying $\mathbf{1} + \mathrm{B} + \mathrm{B}^2 + \mathrm{B}^3$ and $1 - \mathrm{B}$ successively to Y. Show that Z is stationary and express its covariance function using the one of X.
4. Characterize the spectral measure μ of S.
5. Compute the spectral measure of $(1 - \mathrm{B}^4)(Y)$ when X has a spectral density f.

Exercise 3.2 (Canonical ARMA representation). Let $(X_t)_{t\in\mathbb{Z}}$ denote a second-order stationary process satisfying the following recurrence relation

$$X_t - 2X_{t-1} = \varepsilon_t + 4\varepsilon_{t-1}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a second-order white noise with variance σ^2.

1. What is the spectral density of X?
2. What is the canonical representation of X?
3. What is the variance of the innovation process corresponding to X?
4. How is it possible to write X_t as a function of $(\varepsilon_t)_{t\in\mathbb{Z}}$?

Exercise 3.3 (Sum of MA processes). Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ denote two uncorrelated MA processes such that

$$\begin{aligned} X_t &= \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q} \\ Y_t &= \eta_t + \rho_1\eta_{t-1} + ... + \rho_p\eta_{t-p} \end{aligned}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ and $(\eta_t)_{t\in\mathbb{Z}}$ are white noise processes with variance, respectively, σ_ε^2 and σ_η^2. Define $(Z_t)_{t\in\mathbb{Z}}$ by

$$Z_t = X_t + Y_t\,.$$

1. Show that Z is an ARMA process.
2. Assuming that $q = p = 1$ and $0 < \theta_1, \rho_1 < 1$, compute the variance of the innovation process corresponding to Z.

Exercise 3.4 (Sum of AR processes). Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ denote two uncorrelated AR(1) processes :

$$\begin{aligned} X_t &= aX_{t-1} + \varepsilon_t \\ Y_t &= bY_{t-1} + \eta_t \end{aligned}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ and $(\eta_t)_{t\in\mathbb{Z}}$ are white noises variances σ_ε^2 and σ_η^2, respectively, and $0 < a, b < 1$. define Z by

$$Z_t = X_t + Y_t\,.$$

1. Show that there exists a white noise $(\xi_t)_{t\in\mathbb{Z}}$ with variance σ^2 and θ with $|\theta| < 1$ such that
$$Z_t - (a+b)\,Z_{t-1} + abZ_{t-2} = \xi_t - \theta\xi_{t-1}\,.$$
2. Check that
$$\xi_t = \varepsilon_t + (\theta - b)\sum_{k=0}^{\infty}\theta^k\varepsilon_{t-1-k} + \eta_t + (\theta - a)\sum_{k=0}^{\infty}\theta^k\eta_{t-1-k}$$
3. Determine the best linear predictor of Z_{t+1} when (X_s) and (Y_s) are known up to time $s = t$.
4. Determine the best linear predictor of Z_{t+1} when (Z_s) is known up to time $s = t$.
5. Compare the variances of the prediction errors corresponding to the two predictors defined above.

Exercise 3.5. The goal of this exercise is to show that any spectral density f that is continuous on $]-\pi, \pi]$ can be approximated by the spectral density of a moving average process (MA(q)) that equals $|\Theta(\mathrm{e}^{-\mathrm{i}\omega})|^2$ where

$$\Theta\,(\mathrm{B}) = \theta_0 + \theta_1\,\mathrm{B} + \theta_2\,\mathrm{B}^2 + \cdots + \theta_q\,\mathrm{B}^q\ .$$

Let us define $e_k\,(\omega) = \mathrm{e}^{\mathrm{i}k\omega}$ and, for all $n \geq 1$,

$$K_n = \frac{1}{2\pi n}\sum_{j=0}^{n-1}\sum_{k=-j}^{j} e_k\ .$$

1. Compute the integral of K_n over a period.
2. Show that K_n is non-negative and satisfies, for all $\epsilon > 0$, $\sup_{\epsilon\leq|t|\leq\pi} K_n(t) = O\left(n^{-1}\right)$.
3. Deduce that for any continuous (2π)-periodic function g, denoting by
$$g_j\,(\omega) = \sum_{k=-j}^{j} c_k e_k\,(\omega)\ ,$$
its Fourier approximation of order j, where $c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} g\,(\omega)\,\mathrm{e}^{-\mathrm{i}k\omega}d\omega$, then the Cesaro mean $\frac{1}{n}\sum_{j=0}^{n-1} g_j$ converges to g uniformly on $[-\pi, \pi]$.

4. Using this result, show that for all $\varepsilon > 0$, there exists Θ of finite order q such that $\sup_{\omega\in[-\pi,\pi]} \left||\Theta(\mathrm{e}^{-\mathrm{i}\omega})|^2 - f(\omega)\right| < \epsilon$. Suppose first that f is bounded from below by $m > 0$ on $[-\pi, \pi]$.

Exercise 3.6. Let P and Q be defined as in Lemma 3.2.3. Suppose first that $Q(z) = 1 - \alpha z$ for some $\alpha \in \mathbb{C}$.

1. Suppose that $|\alpha| < 1$. Compute δ_1 and δ_2 in this case and exhibit $(\psi_t)_{t\in\mathbb{Z}}$ so that (3.11) holds. What is the value of ψ_t for $t \leq -1$?

2. Do the same when $|\alpha| > 1$. [Hint : use that $Q(z) = -\alpha z(1 - \alpha^{-1}z^{-1})$].

3. Using the partial fraction decomposition of P/Q, prove that Lemma 3.2.3 holds in the general case, leaving aside only the proof of two following assertions:

   (A-1) $\psi_t = 0$ for all $t < 0$ implies $\delta_1 = 0$.

   (A-2) $\psi_t = 0$ for all $t > \deg(P) - \deg(Q)$ implies $\delta_2 = \infty$.

4. Suppose that $\psi_t = 0$ for all $t < 0$. Show that (3.11) implies

$$P(z) = Q(z) \sum_{t\in\mathsf{Z}} \psi_t z^t$$

   for all z such that $|z| < 1$. Deduce that $\delta_1 = 0$. [Hint : use that, by assumption, P and Q do not have common roots.]

5. Use a similar reasoning to prove Assertion *(A-2)*.

Exercise 3.7. Consider the assumptions of Theorem 3.4.2. Express the variance of the white noise of the canonical representation using Φ, Θ and σ^2 (the variance of Z).

Chapter 4

Statistical inference

4.1 Convergence of vector valued random variables

So far we have essentially worked with the L^2 convergence of random variables. Here we recall some standard results for random variables valued in a finite dimensional space $\mathbb{R}^p$ endowed with an arbitrary norm, say the Euclidean norm (denoted by $|x|$). We will use the same definitions and the same notation in this setting as in Appendix C where we have gathered the main useful results about convergence of random variables in general metric spaces. Most of these result should already be known to the reader, perhaps slightly differently expressed. For instance Assertion *(*v*)* in Theorem C.1.8 is usually referred to as *Slutsky's lemma* and is sometimes stated in the following simplest (and less general) form.

Lemma 4.1.1 (Slutsky's Lemma)**.** *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ and $(\mathbf{Y}_n)_{n\in\mathbb{N}}$ be sequences of random variables valued in $\mathbb{R}^p$, $(\mathbf{R}_n)_{n\in\mathbb{N}}$ be sequences of random variables valued in $\mathbb{R}^{q\times p}$, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $\mathbf{X}_n \Longrightarrow \mathbf{X}$, $\mathbf{Y}_n \xrightarrow{P} \mathbf{y}$ and $\mathbf{R}_n \xrightarrow{P} \mathbf{r}$, where $\mathbf{X}$ is a r.v. valued in $rset^p$, $\mathbf{y} \in \mathbb{R}^p$ and $\mathbf{r} \in \mathbb{R}^{q\times p}$. Then we have*

(i) $\mathbf{X}_n + \mathbf{Y}_n \Longrightarrow \mathbf{X} + \mathbf{y}$;

(ii) $\mathbf{R}_n\, \mathbf{X}_n \Longrightarrow \mathbf{r}\, \mathbf{X}$;

(iii) if $\mathbf{r}$ is invertible, $\mathbf{R}_n^{-1}\mathbf{X}_n \Longrightarrow \mathbf{r}^{-1}\mathbf{X}$.

Another example of extensively used result for sequences of vector valued random variables which holds in general metric spaces is the continuous mapping theorem stated as in Theorem C.1.5 (for the weak convergence) or as in Theorem C.1.7 (for the three convergences: strong, in probability and weak).

In contrast, the two following results are specific to the vector valued case, see Jacod and Protter [2003]. The first one indicates how to relate the weak convergence in $\mathbb{R}^p$ to the case $p = 1$.

Theorem 4.1.2 (Cramér-Wold device)**.** *Let $\mathbf{X}, (\mathbf{X}_n)_{n\in\mathbb{N}}$ be random variables valued in $\mathbb{R}^p$ and defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have $\mathbf{X}_n \Longrightarrow \mathbf{X}$ if and only if, for all $\mathbf{t} \in \mathbb{R}^p$, $\mathbf{t}^T\mathbf{X}_n \Longrightarrow \mathbf{t}^T\mathbf{X}$.*

The second result is a characterization of weak convergence using the characteristic functions.

Theorem 4.1.3 (Lévy's theorem)**.** *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of random variables valued in $\mathbb{R}^p$. Denote by ϕ_n the characteristic function of $\mathbf{X}_n$, that is,*

$$\phi_n(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\mathbf{t}^T\mathbf{X}_n}\right], \quad \mathbf{t} \in \mathbb{R}^p .$$

Suppose that $\phi_n(x)$ converges to $\phi(x)$ for all $x \in \mathbb{R}^p$, where ϕ is continuous at the origin. Then there exists a random variable $\mathbf{X}$ valued in $\mathbb{R}^p$ such that $\mathbf{X}$ has characteristic function ϕ and $\mathbf{X}_n \Longrightarrow \mathbf{X}$.

An elementary consequence of this result is the following application to a sequence of Gaussian random variables.

Proposition 4.1.4. *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of Gaussian random p-dimensional vectors. Then the two following assertions are equivalent.*

(i) $\lim_{k\to\infty} \mathbb{E}[\mathbf{X}_k] = \boldsymbol{\mu}$ *and* $\lim_{k\to\infty} \mathrm{Cov}(\mathbf{X}_k) = \Sigma$

(ii) $\mathbf{X}_k \Longrightarrow \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Most of the statistics used for estimation can be written using the empirical measure defined from a set of observations as follows.

Definition 4.1.1 (Empirical measure)**.** *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a sample of n observations in $\mathbb{R}^p$. The empirical measure P_n of $\mathbf{X}_{1:n}$ is the measure on $\mathbb{R}^p$ defined, for all $A \in \mathcal{B}(\mathbb{R}^p)$, by*

$$P_n(A) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_A(\mathbf{X}_k) .$$

For a probability measure P, it is convenient to use the notation $P(h)$ for the expectation $\int h \, \mathrm{d}P$. For instance, following Definition 4.1.1, we will use the notation

$$P_n(h) = \frac{1}{n}\sum_{k=1}^{n} h(\mathbf{X}_k) .$$

The two following classical results apply to i.id. sequences and provide the asymptotic behavior of the empirical measure, see Jacod and Protter [2003].

Theorem 4.1.5 (Law of large numbers and central limit theorem)**.** *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables valued in $\mathbb{R}^p$ with marginal distribution P. Then the two following assertions hold.*

- *Law of large numbers : for any measurable $h : \mathbb{R}^p \to \mathbb{R}^q$ such that $P|h| < \infty$, we have*

$$P_n(h) \xrightarrow{\text{a.s.}} P(h) .$$

- *Central limit theorem : for any measurable $h : \mathbb{R}^p \to \mathbb{R}^q$ such that $P(|h|^2) < \infty$, we have*

$$\sqrt{n}\,(P_n(h) - P(h)) \Longrightarrow \mathcal{N}(0, P(hh^T) - P(h)P(h^T)) .$$

An alternative way to prove an a.s. convergence is to rely on the Borel Cantelli lemma, see Lemma C.1.1.

In the setting of vector valued random variables, simple asymptotic results are conveniently expressed using the *stochastic order symbols*.

Definition 4.1.2 (Stochastic order symbols)**.** *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of random variables valued in $\mathbb{R}^p$ and defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will say that $\mathbf{X}_n$ is stochastically negligible and denote $\mathbf{X}_n = o_P(1)$ if $\mathbf{X}_n \xrightarrow{P} 0$ that is, for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\mathbf{X}_n| > \epsilon) = 0 \, .$$

We will say that $\mathbf{X}_n$ is stochastically bounded and denote $\mathbf{X}_n = O_P(1)$ if $(\mathbf{X}_n)_{n\in\mathbb{N}}$ is tight, that is,

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(|\mathbf{X}_n| > M) = 0 \, .$$

Moreover, for a sequence $(R_n)_{n\in\mathbb{N}}$ of random variables valued in $\mathbb{R}_+$ and defined on $(\Omega, \mathcal{F}, \mathbb{P})$, we will write $\mathbf{X}_n = o_P(R_n)$ (resp. $\mathbf{X}_n = O_P(R_n)$) if $\mathbf{X}_n/R_n = o_P(1)$ (resp. $\mathbf{X}_n/R_n = O_P(1)$) with the convention $0/0 = 0$.

The definition of tightness used above for defining the symbol $O_P(1)$ corresponds to the one given in Appendix C for a general set of probability measures defined on a metric space. Namely, $(\mathbf{X}_n)_{n\in\mathbb{N}}$ is tight means that the set of image probability measures $\{\mathbb{P}\circ\mathbf{X}_n^{-1},\, n\in\mathbb{N}\}$ is tight in the sense of Definition C.2.1 as a set of probability measures on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. We have the following result which says how the symbol $O_P(1)$ is related to the weak convergence.

Theorem 4.1.6. *Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of random variables valued in $\mathbb{R}^p$. Then the two following assertions hold.*

(i) If $\mathbf{X}_n$ converges weakly, then $\mathbf{X}_n = O_P(1)$.

(ii) If $\mathbf{X}_n = O_P(1)$, then there exists a subsequence $(\mathbf{X}_{\alpha_n})$ such that $\mathbf{X}_{\alpha_n}$ converges weakly.

Proof. We first prove *(i)*. Suppose that $\mathbf{X}_n \Longrightarrow \mathbf{X}$ for some r.v. $\mathbf{X}$. Then $|\mathbf{X}_n| \Longrightarrow |\mathbf{X}|$, and for any continuity point M of the distribution function of $|\mathbf{X}|$, we have

$$\lim_{n\to\infty} \mathbb{P}(|\mathbf{X}_n| > M) = \mathbb{P}(|\mathbf{X}| > M) \, .$$

Since the set of discontinuity points of the distribution function of $|\mathbf{X}|$ is at most countable, we can let M go to infinity and obtain that $\mathbf{X}_n = O_P(1)$.

To conclude the proof we observe that Theorem C.2.3 implies *(ii)*. □

The Stochastic order symbols o_P and O_P can be used as the deterministic ones, namely, we have the following result, the proof of which is left to the reader (see Exercise 4.1).

Proposition 4.1.7. *The following relations hold for sequences of vector valued random variables with compatible dimensions.*

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1), \\ O_P(1) + O_P(1) &= O_P(1), \\ O_P(1) \times o_P(1) &= o_P(1) \, . \end{aligned}$$

Finally we recall another standard result for sequences if vector valued random variables, the so called δ-method, which allows us to obtain the weak convergence of the sequence $r_n(g(\mathbf{X}_n) - g(\mathbf{x}))$ given that of $r_n(\mathbf{X}_n - \mathbf{x})$ under practical conditions.

Proposition 4.1.8 (δ-method)**.** *Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be a measurable function which is differentiable at $\mathbf{x} \in \mathbb{R}^k$. Let $\mathbf{Y}, (\mathbf{X}_n)_{n\in\mathbb{N}}$ be a sequence of random variables valued in $\mathbb{R}^p$ and $(r_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers such that $\lim_n r_n = \infty$. Suppose that $r_n(\mathbf{X}_n - \mathbf{x}) \Longrightarrow \mathbf{Y}$. Then we have $r_n(g(\mathbf{X}_n) - g(\mathbf{x})) \Longrightarrow \partial g(\mathbf{x})^T \mathbf{Y}$.*

Proof. Since g is differentiable at $\mathbf{x}$, we have

$$g(\mathbf{X}_n) - g(\mathbf{x}) = \partial g(\mathbf{x})^T (\mathbf{X}_n - \mathbf{x}) + R(\mathbf{X}_n - \mathbf{x}) \,,$$

where $R(\mathbf{x}) = o(\mathbf{x})$ as $\mathbf{x} \to 0$. Multiplying by r_n we get

$$r_n(g(\mathbf{X}_n) - g(\mathbf{x})) = \partial g(\mathbf{x})^T \left(r_n(\mathbf{X}_n - \mathbf{x})\right) + r_n R(\mathbf{X}_n - \mathbf{x}) \,.$$

Now the first term in the right-hand side converges weakly to $\partial g(\mathbf{x})^T \mathbf{Y}$ by the continuous mapping theorem. Since $r_n \to \infty$, and $r_n(\mathbf{X}_n - \mathbf{x}) = O_p(1)$, we have $\mathbf{X}_n - \mathbf{x} = o_P(1)$ and thus $R(\mathbf{X}_n - \mathbf{x}) = o_P(|\mathbf{X}_n - \mathbf{x}|)$. Hence the second term is $o_P(1)$ and we conclude with Slutsky's Lemma. □

4.2 Empirical estimation of the mean and autocovariance function

Let $p \geq 1$ and $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ be a $\mathbb{C}^p$-valued weakly stationary process with mean $\boldsymbol{\mu}$ (valued in $\mathbb{C}^p$) and autocovariance function Γ (valued in $\mathbb{C}^{p\times p}$). We wish to estimate $\boldsymbol{\mu}$ and Γ based on a finite sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

To this end, we introduce two classical estimators.

Definition 4.2.1. *The* empirical mean *(or* sample mean*) and the* empirical autocovariance function *of the sample $\mathbf{X}_{1:n}$ are respectively defined as*

$$\widehat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_t \tag{4.1}$$

$$\widehat{\Gamma}_n(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H & \text{if } 0 \leq h \leq n-1 \,, \\ n^{-1} \sum_{t=1-h}^{n} (\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H & \text{if } 0 \leq -h \leq n-1 \,, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Remark 4.2.1. *To avoid separating the different cases for h, the right-hand side of (4.2) can be written as follows*

$$\widehat{\Gamma}_n(h) = n^{-1} \sum_{1\leq t, t+h\leq n} (\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H \,, \tag{4.3}$$

where, by convention, the sum is zero if there is no $t \in \mathbb{Z}$ such that $1 \leq t, t+h \leq n$.

It is tempting to replace the normalizing term n^{-1} by $(n-|h|)^{-1}$ in (4.2), at least when $|h|<n$ as $n-|h|$ is the number of terms in the sum. As $n\to\infty$ for a fixed h, the two normalizations are equivalent. We actually prefer the normalizing term n^{-1} because it yields a very interesting property for $\widehat{\Gamma}_n$, namely, it is an autocovariance function. To see why, let us introduce a new statistic of interest.

Definition 4.2.2 (Periodogram)**.** *The periodogram of the sample $\mathbf{X}_{1:n}$ is the function valued in $\mathbb{C}^{p\times p}$ and defined on $\mathbb{T}$ by*

$$\mathbf{I}_n(\lambda)=\frac{1}{2\pi n}\left(\sum_{t=1}^{n}(\mathbf{X}_t-\widehat{\boldsymbol{\mu}}_n)\mathrm{e}^{-\mathrm{i}t\lambda}\right)\left(\sum_{t=1}^{n}(\mathbf{X}_t-\widehat{\boldsymbol{\mu}}_n)\mathrm{e}^{-\mathrm{i}t\lambda}\right)^{H}. \tag{4.4}$$

Then we have the following result.

Theorem 4.2.1. *Let $X_{1:n}$ be a sample of scalar observations. Let $\widehat{\gamma}_n$ and I_n denote its empirical autocovariance function and its periodogram. Then $\widehat{\gamma}_n$ satisfies the properties of Proposition 2.3.1, hence it is an admissible autocovariance function. Moreover I_n is the corresponding spectral density and, either $\widehat{\gamma}_n\equiv 0$ or the matrix $\widehat{\Gamma}^{+}_{n,p}$ is invertible for all $p\geq 1$, where*

$$\widehat{\Gamma}^{+}_{n,p}=\begin{bmatrix}\widehat{\gamma}_n(0) & \widehat{\gamma}_n(-1) & \cdots & \widehat{\gamma}_n(-p+1)\\ \widehat{\gamma}_n(1) & \widehat{\gamma}_n(0) & \cdots & \widehat{\gamma}_n(-p+2)\\ \vdots & & & \\ \widehat{\gamma}_n(p-1) & \widehat{\gamma}_n(p-2) & \cdots & \widehat{\gamma}_n(0)\end{bmatrix}.$$

Remark 4.2.2. *Observe that, for a sample $\mathbf{X}_{1:n}$ of vector observations and for any $\mathbf{t}\in\mathbb{C}^p$, the empirical autocovariance function of $\mathbf{t}^H\mathbf{X}_{1:n}$ and its periodogram are given by $\widehat{\gamma}_n=\mathbf{t}^H\widehat{\Gamma_n}\mathbf{t}$ and $I_n(\lambda)=\mathbf{t}^H\mathbf{I}_n(\lambda)\mathbf{t}$, where $\widehat{\Gamma_n}$ and $\mathbf{I}_n(\lambda)$ are the empirical autocovariance function and the periodogram of $\mathbf{X}_{1:n}$. Hence Theorem 4.2.1 also implies that in the vector case, the empirical autocovariance function is an admissible covariance function.*

Proof of Theorem 4.2.1. Observe that I_n is a nonnegative function. Moreover, we have

$$\begin{aligned}\int_{\mathbb{T}}\mathrm{e}^{\mathrm{i}\lambda h}I_n(\lambda)\mathrm{d}\lambda&=\frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n}(X_s-\widehat{\mu}_n)\overline{(X_t-\widehat{\mu}_n)}\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}\lambda(h-s+t)}\\&=\widehat{\gamma}_n(h)\,,\end{aligned}$$

since

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}\lambda(h-s+t)}=\begin{cases}1 & \text{if } s=h+t,\\ 0 & \text{otherwise.}\end{cases}$$

By Theorem 2.4.1, we get that $\widehat{\gamma}_n$ is a nonnegative hermitian function.

Consider now two cases. First, if $\widehat{\gamma}_n(0)=0$, then $\widehat{\gamma}_n\equiv 0$ (since $\widehat{\gamma}_n$ is an admissible covariance function). Second, if $\widehat{\gamma}_n(0)>0$, since $\widehat{\gamma}_n(h)\to\infty$ as $h\to\infty$, Proposition 2.4.3 implies that $\widehat{\Gamma}^{+}_{n,p}$ is invertible for all $p\geq 1$. □

Since $\widehat{\mu}_n$ is linear with respect to the observations and the first and second order moments of a weakly stationary process is given by its mean and autocovariance function, it is easy to derive the first and second order behavior of $\widehat{\mu}_n$ in this case. This is done in the following result in the real valued case for simplicity.

Theorem 4.2.2. *Let (X_t) be a real-valued weakly stationary process with mean μ and autocovariance function γ. Let $\widehat{\mu}_n$ denote the empirical mean of the sample $X_{1:n}$. Then the following assertions hold.*

(i) $\widehat{\mu}_n$ *is an* unbiased estimator *of* μ*, that is,* $\mathbb{E}[\widehat{\mu}_n] = \mu$ *for all* $n \geq 1$.

(ii) If $\lim_{h\to\infty} \gamma(h) = 0$*, then* $\lim_{n\to\infty} \mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right] = 0$.

(iii) If moreover $\gamma \in \ell^1$*, then, as* $n \to \infty$*,*

$$\operatorname{Var}(\widehat{\mu}_n) \leq n^{-1} \|\gamma\|_1 , \tag{4.5}$$

$$\operatorname{Var}(\widehat{\mu}_n) = n^{-1}(2\pi f(0) + o(1)) , \tag{4.6}$$

where f is the spectral density of (X_t).

Proof. Assertion *(i)* is immediate and implies that $\mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right] = \operatorname{Var}(\widehat{\mu}_n)$. Thus we have

$$\begin{aligned}\operatorname{Var}(\widehat{\mu}_n) &= n^{-2} \sum_{s=1}^{n} \sum_{t=1}^{n} \operatorname{Cov}(X_s, X_t) \\ &= n^{-1} \sum_{\tau\in\mathbb{Z}} (1 - |\tau|/n)_+ \gamma(\tau) ,\end{aligned}$$

where we set $\tau = s - t$ and used the notation $a_+ = \max(a, 0)$. From this expression, we easily get Assertion *(ii)* and (4.5). Under the assumption of *(iii)*, we may apply the dominated convergence and get that

$$\lim_{n\to\infty} n \operatorname{Var}(\widehat{\mu}_n) = \sum_{\tau=-\infty}^{\infty} \lim_{n\to\infty} (1 - |\tau|/n)\, \gamma(\tau) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) = 2\pi f(0) .$$

Hence we have (4.6). □

4.3 Consistency of the empirical mean and of the empirical autocovariance function

We now investigate some simple conditions under which the empirical mean $\widehat{\boldsymbol{\mu}}_n$ and the empirical autocovariance function $\widehat{\Gamma}_n$ are consistent estimators of $\boldsymbol{\mu}$ and Γ, that is, $\widehat{\boldsymbol{\mu}}_n$ converges to $\boldsymbol{\mu}$ and $\widehat{\Gamma}_n(h)$ converges to $\Gamma(h)$ for all $h \in \mathbb{Z}$ as $n \to \infty$. If the convergence holds a.s. we shall say that the estimator is *strongly consistent* and if it holds in probability, we shall say that the estimator is *weakly consistent*.

We recalled the law of large numbers in Theorem 4.1.5, which states that in the i.i.d. case, the empirical mean is a strongly consistent estimator of the mean, that is, $\widehat{\boldsymbol{\mu}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}$ as $n \to \infty$. The ergodic theory provides a generalization of this results to a much general class of strongly stationary processes, namely the class of *ergodic processes*. However in these lecture notes, we shall consider a more elementary approach to the consistency. More precisely, it is in general easier to find sufficient conditions for an L^2 convergence by controlling the bias and the variance and it directly implies the weak convergence by the Markov inequality. Some refinement of this approach further allows to obtain the strong consistency, using the Borel Cantelli lemma.

Theorem 4.3.1. *Let (X_t) be a real-valued weakly stationary process with mean μ and autocovariance function γ. Let $\widehat{\mu}_n$ denote the empirical mean of the sample $X_{1:n}$, seen as an estimator of μ. Then the following assertions hold.*

(i) If $\lim_{h\to\infty}\gamma(h)=0$, the estimator $\widehat{\mu}_n$ is weakly consistent.

(ii) If moreover $\gamma\in\ell^1$, then the estimator $\widehat{\mu}_n$ is strongly consistent.

Proof. By the Markov inequality, the convergence in probability is a consequence of the L^2 convergence. Thus, Assertion *(i)* directly follows from Assertion *(ii)* in Theorem 4.2.2.

Let us now prove Assertion *(ii)*. First, (4.6) and the Markov inequality imply that, for all $\epsilon>0$,

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|\widehat{\mu}_{n^2}-\mu|\geq\epsilon\right)<\infty\,.$$

By Lemma C.1.1, we get $\lim_{n\to\infty}\widehat{\mu}_{n^2}=0$ a.s. We now need to extend this result to the sequence $(\widehat{\mu}_n)$. We write

$$\widehat{\mu}_n-\mu=\frac{m_n}{n}(\widehat{\mu}_{m_n}-\mu)+n^{-1}\sum_{s=m_n+1}^{n}(X_s-\mu)\,, \tag{4.7}$$

with $m_n=\lfloor\sqrt{n}\rfloor^2$. Since m_n/n is bounded, we already know the first term in the right-hand side converges to 0 a.s. The second term is centered and has the same variance as $n^{-1}\widehat{\mu}_{n-m_n}$, hence of order $O((n-m_n)/n^2)=O((n-m_n)/n^2)=O(n^{1/2-2})$. Proceeding as above, Lemma C.1.1 yields that the second term in the right-hand side of (4.7) also converges to 0 a.s. This concludes the proof. □

The weak consistency amounts to saying that the *confidence interval* $[\widehat{\mu}_n-\epsilon,\widehat{\mu}_n+\epsilon]$ contains the true parameter μ with probability tending to 1 as the number of observations $n\to\infty$. Thanks to this simple statistical application and because it is easier to prove than strong consistency, we shall mainly use this type of consistency in the following, in particular when considering covariance estimation.

Also observe that we stated Theorem 4.3.1 in the case of a real-valued process. Since for both the a.s. convergence and the convergence in probability, the convergence of a vector is equivalent to the convergence of its components, it follows that the same result also holds in the case of a $\mathbb{C}^p$-valued process.

Similarly we shall provide sufficient conditions for the weak consistency of the empirical autocovariance function in the case of a real-valued process. The multi-dimensional case then follows by writing, for two real-valued processes (X_t) and (Y_t),

$$\mathrm{Cov}\,(X_s,Y_t)=\frac{1}{2}\left[\mathrm{Cov}\,(X_s+Y_s,X_t+Y_t)-\mathrm{Cov}\,(X_s-Y_s,X_t-Y_t)\right]\,,$$

and by observing that a similar relation holds for the corresponding empirical covariances. Hence applying a consistency result to the real-valued processes (X_t+Y_t) and (X_t-Y_t) implies a consistency result which applies to the cross-covariance function $\mathrm{Cov}\,(X_s,Y_t)$.

To obtain a weak consistency result on the empirical autocovariance function, we shall rely on the computation of its mean and variance. Since this variance requires the expectation of the product of 4 r.v. X_s (moments of 4th order of the process X), the second order

properties of the underlying process is no longer sufficient to carry out the computation. Hence additional conditions are necessary. The simplest one is to assume that X is Gaussian but it is very restrictive in practice. Instead we shall use a linear representation of X, say the following assumption.

Assumption 4.3.1. *$X = (X_t)_{t\in\mathbb{Z}}$ is a real valued* linear process with short memory*, that is, it admits the representation*

$$X = \mu + \mathrm{F}_\psi(Z) \,, \tag{4.8}$$

where $\mu \in \mathbb{R}$, $Z \sim \mathrm{WN}(0, \sigma^2)$ is real valued and $(\psi_t)_{t\in\mathbb{Z}} \in \ell^1$ is also real valued.

Then, by Corollary 3.1.3, X is a weakly stationary process with mean μ, autocovariance function γ and spectral density function f given by

$$\gamma(h) = \sigma^2 \sum_{k\in\mathbb{Z}} \psi_{k+h}\psi_k \tag{4.9}$$

$$f(\lambda) = \frac{1}{2\pi} \sum_{\tau\in\mathbb{Z}} \gamma(\tau)\mathrm{e}^{-\mathrm{i}\tau\lambda} \,. \tag{4.10}$$

Now, to compute the 4th order moments of X, we shall just need an assumption on Z. We shall use the following one.

Assumption 4.3.2. *The centered white noise Z satisfies, for a constant $\eta \geq 1$, for all $s \leq t \leq u \leq v$,*

$$\mathbb{E}[Z_s\, Z_t\, Z_u\, Z_v] = \begin{cases} \eta\sigma^4 & \text{if } s = t = u = v, \\ \sigma^4 & \text{if } s = t < u = v, \\ 0 & \text{otherwise.} \end{cases}$$

A simple example is the case where Z is a strong white noise with finite 4th order moment. More generally Assumption 4.3.2 holds if $\mathbb{E}[Z_t^4] = \eta\sigma^4$, $\mathbb{E}[Z_t^3|\mathcal{F}_{t-1}] = \mathbb{E}[Z_t|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[Z_t^2|\mathcal{F}_{t-1}] = \sigma^2$ for all $t \in \mathbb{Z}$, where $\mathcal{F}_t$ is a filtration with respect to which Z is adapted (Z_s is $\mathcal{F}_t$-measurable for all $s \leq t$).

A direct consequence is the following lemma, whose proof is left to the reader (see Exercise 4.2).

Lemma 4.3.2. *Suppose that Assumption 4.3.1 and Assumption 4.3.2 hold with $\mu = 0$. Then, for all $k, l, p, q \in \mathbb{Z}$*

$$\begin{aligned}\mathbb{E}\left[X_k X_\ell X_p X_q\right] = (\eta - 3)\sigma^4 \sum_{i\in\mathbb{Z}} \psi_{k+i}\psi_{\ell+i}\psi_{p+i}\psi_{q+i} + \gamma(k-\ell)\gamma(p-q) \\ + \gamma(k-p)\gamma(\ell-q) + \gamma(k-q)\gamma(\ell-p) \,,\end{aligned} \tag{4.11}$$

where γ is the autocovariance function of X. Moreover, there exists a constant C such that, for all $m \in \mathbb{N}$,

$$\mathbb{E}\left[\left(\sum_{t=1}^{m} X_t\right)^4\right] \leq Cm^2 \,. \tag{4.12}$$

Let us now state a weak consistency result for the empirical covariance function of a real valued process.

Theorem 4.3.3. *Suppose that Assumption 4.3.1 and Assumption 4.3.2 hold. Let $\widehat{\gamma}_n$ denote the empirical autocovariance function of the sample $X_{1:n}$. Then, for all $p, q \in \mathbb{Z}$,*

$$\mathbb{E}\left[\widehat{\gamma}_n(p)\right] = \gamma(p) + O(n^{-1}) \,, \tag{4.13}$$

$$\lim_{n\to\infty} n \operatorname{Cov}\left(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q)\right) = V(p,q) \,, \tag{4.14}$$

where γ is the autocovariance function of X and

$$V(p,q) = (\eta - 3)\gamma(p)\gamma(q) + \sum_{u\in\mathbb{Z}} \left[\gamma(u)\gamma(u-p+q) + \gamma(u+q)\gamma(u-p)\right] \,. \tag{4.15}$$

In particular, $\widehat{\gamma}_n(p)$ is a weakly consistent estimator of $\gamma(p)$,

$$\widehat{\gamma}_n(p) = \gamma(p) + O_P(n^{-1/2}) \,. \tag{4.16}$$

Proof. We first observe that replacing X by $X - \mu$ does not modify the definitions of $\widehat{\gamma}_n$ and γ. Hence we can set $\mu = 0$ without loss of generality.

The Markov inequality, (4.13) and (4.14) yield (4.16). Hence it only remains to prove (4.13) and (4.14).

To this end, we introduce

$$\tilde{\gamma}_n(h) = n^{-1} \sum_{t=1}^{n} X_{t+h} X_t \,. \tag{4.17}$$

This is an unbiased estimator of $\gamma(h)$ when X is known to be centered, which we assumed in this proof. However it is different from $\widehat{\gamma}_n(h)$ even in the centered case. First this estimator uses more observations (since $t + h$ is not required to be in $\{1, \dots, n\}$), and second $\hat{\mu}_n$ does not vanish, even in the centered case. More precisely, we have, for all $h \in \mathbb{Z}$,

$$\widehat{\gamma}_n(h) - \tilde{\gamma}_n(h) = -\sum_{t\in\Delta_{n,h}} (X_{t+h} - \hat{\mu}_n)(X_t - \hat{\mu}_n) - \hat{\mu}_n \left[\hat{\mu}_n - \frac{1}{n}\sum_{t=1}^{n}(X_{t+h} + X_t)\right] \,,$$

where $\Delta_{n,h} = \{1, \dots, n\} \setminus \{1-h, \dots, n-h\}$ has cardinality at most $|h|$. By Lemma 4.3.2, we have $\mathbb{E}\left[(\hat{\mu}_n)^4\right] = O(n^{-2})$ and

$$\mathbb{E}\left[\left(\sum_{t=1}^{n} X_{t+h}\right)^4\right] = \mathbb{E}\left[\left(\sum_{t=1}^{n} X_t\right)^4\right] = O(n^2) \,.$$

Thus, by the Cauchy-Schwarz inequality we get that, for all $h \in \mathbb{Z}$,

$$\mathbb{E}\left[\left(\widehat{\gamma}_n(h) - \tilde{\gamma}_n(h)\right)^2\right] = O(n^{-2}) \tag{4.18}$$

By Jensen's inequality, we get

$$\mathbb{E}\left[\widehat{\gamma}_n(p)\right] = \mathbb{E}\left[\tilde{\gamma}_n(p)\right] + O(n^{-1}) \,, \tag{4.19}$$

Since $\tilde{\gamma}_n(p)$ is an unbiased estimator of $\gamma(p)$, this yields (4.13). Next, we have, for all $p, q \in \mathbb{Z}$,

$$\operatorname{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) = n^{-2} \sum_{s=1}^{n}\sum_{t=1}^{n} \operatorname{Cov}\left(X_{s+p}X_s, X_{t+q}X_t\right) \,.$$

By Lemma 4.3.2, we know that

$$\begin{aligned}\mathrm{Cov}\left(X_{s+p}X_s, X_{t+q}X_t\right) = (\eta-3)\sigma^4 \sum_{i\in\mathbb{Z}} \psi_{s+i}\psi_{s+p+i}\psi_{t+i}\psi_{t+q+i} \\ + \gamma(s-t+p-q)\gamma(s-t) + \gamma(s+p-t)\gamma(s-t-q)\ .\end{aligned}$$

Note that this term is unchanged when shifting s and t by the same constant. Hence it can be written as $v(s-t)$ where

$$v(u) = (\eta-3)\sigma^4 \sum_{i\in\mathbb{Z}} \psi_{u+i}\psi_{u+p+i}\psi_i\psi_{q+i} + \gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q)\ . \tag{4.20}$$

Hence we get, for all $p, q \in \mathbb{Z}$,

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) &= n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n} v(s-t) \\ &= n^{-2}\sum_{\tau\in\mathbb{Z}} (n-|\tau|)_+\, v(\tau)\ .\end{aligned}$$

Using that $\psi \in \ell^1$, we easily get that γ and v are in ℓ^1. It follows that as $n \to \infty$,

$$\mathrm{Cov}\left(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q)\right) \sim n^{-1}\sum_{\tau\in\mathbb{Z}} v(\tau)\ .$$

Now, by (4.9) and (4.15), we have

$$\sum_{\tau\in\mathbb{Z}} v(\tau) = V(p,q)\ .$$

Hence we get that, as $n \to \infty$,

$$\mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) \sim n^{-1}V(p,q)\ . \tag{4.21}$$

From this and (4.18), it follows that, for all $p, q \in \mathbb{Z}$,

$$\mathrm{Cov}\left(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q)\right) = \mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) + O(n^{-3/2})\ . \tag{4.22}$$

Finally, (4.21) and (4.22) yield (4.14), which concludes the proof. □

4.4 Asymptotic distribution of the empirical mean

From Theorem 4.2.2, we know that, under suitable assumptions, $\hat{\mu}_n$ is an unbiased estimator with variance asymptotically behaving as $O(n^{-1})$. A natural question is thus to determine the convergence of $\sqrt{n}(\hat{\mu}_n - \mu)$. This is useful to build confidence intervals for the mean with given asymptotic confidence level. In the i.i.d. case, the central limit Theorem (see Theorem 4.1.5) indicates that this sequence converge weakly to a Gaussian distribution. We may hope that such a result extends for more general time series. As in Theorem 4.3.3, we need to somehow precise the distribution of the time series by relying on Assumption 4.3.1 with a suitable assumption on the white noise Z, namely.

Assumption 4.4.1. *The centered white noise* $(Z_t)_{t\in\mathbb{Z}}$ *satisfies*

$$n^{-1/2}\sum_{t=1}^{n} Z_t \Longrightarrow \mathcal{N}(0,\sigma^2)\ .$$

A first generalization of the CLT in Theorem 4.1.5 is given by the following result.

Proposition 4.4.1. *Suppose that Assumption 4.3.1 and Assumption 4.4.1 hold with a finitely supported sequence* ψ*. Let* $\widehat{\mu}_n$ *denote the empirical mean of the sample* $X_{1:n}$*. Then, as* $n\to\infty$*,*

$$\sqrt{n}\,(\widehat{\mu}_n-\mu)\Longrightarrow\mathcal{N}(0,2\pi f(0)) \tag{4.23}$$

where $f(\lambda)=\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma(\tau)\mathrm{e}^{-\mathrm{i}\tau\lambda}$ *is the spectral density of* (X_t)*.*

Proof. Since $\widehat{\mu}_n-\mu$ is the empirical mean of the sample $\bar{X}_{1:n}$ with $\bar{X}_k=X-\mu$, we can assume $\mu=0$ by replacing X by $X-\mu$.

Let m be such that $[-m,m]$ contains the support of ψ. Denote $\widehat{\mu}_n^Z=n^{-1}\sum_{t=1}^{n}Z_t$. Then we have

$$\begin{aligned}\widehat{\mu}_n&=\sum_{j=-m}^{m}\psi_j\left(n^{-1}\sum_{t=1}^{n}Z_{t-j}\right)\\&=\left(\sum_{j\in\mathbb{Z}}\psi_j\right)\widehat{\mu}_n^Z+n^{-1}\sum_{j=-m}^{m}\psi_j R_{n,j}\ ,\end{aligned}$$

where $|R_{n,j}|\le\sum_{s\in I_{n,j}}|Z_s|$ and $I_{n,j}$ is the symmetric difference between $\{1,\dots,n\}$ and $\{1-j,\dots,n-j\}$ (that is, the set of indices that are in one and only one of these two sets). Since the cardinality of $I_{n,j}$ is at most $2j$, we have $(\mathbb{E}|R_{n,j}|^2)^{1/2}\le 2j\sigma$ and thus $\sum_{j=-m}^{m}\psi_j R_{n,j}=O_p(1)$. Hence we obtain (4.23). □

Now, we can state a more general extension similar to Proposition 4.4.1 but without the assumption on the support of ψ. The idea is to approximate $X=\mathrm{F}_\psi(Z)$ by $\mathrm{F}_{\psi^m}(Z)$ where ψ_m has a finite support.

Theorem 4.4.2. *Suppose that Assumptions 4.3.1 and 4.4.1 hold. Let* $\widehat{\mu}_n$ *denote the empirical mean of the sample* $X_{1:n}$*. Then the CLT (4.23) holds.*

Proof. As in the proof of Proposition 4.4.1, we can assume that $\mu=0$ without loss of generality.

Define the sequence ψ^m by

$$\psi_k^m=\begin{cases}\psi_k & \text{if } |k|\le m,\\ 0 & \text{otherwise.}\end{cases} \tag{4.24}$$

Let $\widehat{\mu}_n^m$ be the empirical mean of the the sample $[\mathrm{F}_{\psi^m}(Z)]_{1:n}$. Then by Proposition 4.4.1, we have, for all $m\ge 1$, as $n\to\infty$,

$$\sqrt{n}\,(\widehat{\mu}_n^m-\mu)\Longrightarrow\mathcal{N}(0,\sigma_m^2)\ , \tag{4.25}$$

where

$$\sigma_m^2 = \left(\sum_{j=-m}^{m} \psi_j \right)^2 .$$

Moreover, using that $\psi \in \ell^1$, we have $\sigma_m^2 \to 2\pi f(0)$ as $m \to \infty$ and thus, applying Proposition 4.1.4, as $n \to \infty$,

$$\mathcal{N}(0, \sigma_m^2) \Longrightarrow \mathcal{N}(0, 2\pi f(0)) .$$

By Lemma C.1.9, this convergence and (4.25) imply (4.23) if for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}|\widehat{\mu}_n - \widehat{\mu}_n^m| > \epsilon) = 0 . \tag{4.26}$$

Hence it only remains to show (4.26). Observe that, by linearity of the empirical mean, $\widehat{\mu}_n - \widehat{\mu}_n^m$ is the empirical mean of $[\mathrm{F}_{\psi-\psi^m}(Z)]_{1:n}$. Moreover the process $\mathrm{F}_{\psi-\psi^m}(Z)$ has an autocovariance function γ_n with ℓ^1-norm satisfying $\|\gamma_n\|_1 \leq \left(\sum_{|j|>m} |\psi_j|\right)^2$. Applying Theorem 4.2.2, we get

$$\mathbb{E}\left[(\widehat{\mu}_n - \widehat{\mu}_n^m)^2\right] = \mathrm{Var}\,(\widehat{\mu}_n - \widehat{\mu}_n^m) \leq n^{-1} \left(\sum_{|j|>m} |\psi_j| \right)^2 .$$

Assertion (4.26) follows by the Markov inequality. □

Proposition 4.4.1 and Theorem 4.4.2 heavily rely on the linear representation of Assumption 4.3.1. It is interesting to note that the above technique can be applied in the following framework which do not assume a linear representation.

Definition 4.4.1. *Le $m \geq 1$. A process $X = (X_t)_{t\in\mathbb{Z}}$ is said to be* m-dependent *if, for all $t \in \mathbb{Z}$, $(X_s)_{s\leq t}$ and $(X_s)_{s>t+m}$ are independent.*

Theorem 4.4.3. *Let X be an L^2 real valued strictly stationary* m-dependant *process with mean μ and autocovariance function γ. Let $\widehat{\mu}_n$ denote the empirical mean of the sample $X_{1:n}$. Then the CLT (4.23) holds.*

Proof. As usual, we can assume $\mu = 0$ without loss of generality.

The proof relies on an approximation of $\hat{\mu}_n$ by weakly convergent sequences (denoted by $\hat{\mu}_n^p$ below) and then by making use of Lemma C.1.9, as in the proof of Theorem 4.4.2. Let $p \geq 1$ and define the integers p and r by the Euclidean division $n = (p+m)k + r$. Then we have

$$\begin{aligned}
\hat{\mu}_n &= n^{-1} \sum_{j=1}^{k-1} \sum_{s=1}^{p+m} X_{j(p+m)+s} + n^{-1} \sum_{s=1}^{r} X_{k(p+m)+s} \\
&= n^{-1} \sum_{j=1}^{k-1} S_{j,p} + R_{n,p} ,
\end{aligned}$$

where we defined

$$S_{j,p} = \sum_{s=1}^{p} X_{j(p+m)+s}$$

and

$$R_{n,p} = n^{-1} \sum_{j=1}^{k-1} \sum_{s=p+1}^{p+m} X_{j(p+m)+s} + n^{-1} \sum_{s=1}^{r} X_{k(p+m)+s} \ .$$

Using that $(X_t)_{t\in\mathbb{Z}}$ is m-dependent, we get that $(S_{p,j})_{j\geq 1}$ is an i.i.d. sequence. Hence the CLT in Theorem 4.1.5 applies and we have

$$n^{-1/2} \sum_{j=1}^{k-1} S_{j,p} \Longrightarrow \mathcal{N}(0, \sigma_p^2) \ ,$$

where

$$\sigma_p^2 = \mathrm{Var}(S_{1,p}) = \sum_{\tau\in\mathbb{Z}} (p - |\tau|)_+ \gamma(\tau) \ .$$

Observe that σ_p^2 converges to $2\pi f(0)$ as $p \to \infty$. So, by Lemma C.1.9, to conclude the proof, it only remains to show that, for all $\epsilon > 0$,

$$\lim_{p\to\infty} \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}|R_{n,p}| > \epsilon) = 0 \ . \tag{4.27}$$

We first observe that, since $r \leq p + m$ for all n, we have

$$\mathrm{Var}\left(\sum_{s=1}^{r} X_{k(p+m)+s} \right) \leq C\,(p+m) \ , \tag{4.28}$$

where

$$C = \sum_{\tau\in\mathbb{Z}} |\gamma(\tau)| \ .$$

Now, for $p \geq m$ the sums $\sum_{s=p+1}^{p+m} X_{j(p+m)+s}$, $j \geq 1$ are i.i.d., and we find that

$$\mathrm{Var}\left(\sum_{j=1}^{k-1} \sum_{s=p+1}^{p+m} X_{j(p+m)+s} \right) \leq C\,(k-1)m \ .$$

Hence, with (4.28) and the definition of $R_{n,p}$ above, we get that, for $p \geq m$,

$$\mathbb{E}\left[nR_{n,p}^2 \right] \leq 2n^{-1} C\,(km + p) \leq 2C\,((m+p)^{-1} + pn^{-1}) \ .$$

Hence, by the Markov inequality, we obtain (4.27) and the proof is finished. □

4.5 Asymptotic distribution of the empirical autocovariance function

An approach similar to Theorem 4.4.2 can be used to derive the asymptotic distribution of the empirical autocovariance function under a more precise assumption on the white noise Z of the linear representation (4.8).

Assumption 4.5.1. *The centered white noise $(Z_t)_{t\in\mathbb{Z}}$ is a strong noise and $\mathbb{E}[Z_0^4] = \eta\sigma^4$ for some $\eta \geq 1$.*

We already mentioned that this assumption implies Assumption 4.3.2. Thus the asymptotic behavior of the covariances $\mathrm{Cov}\,(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q))$ are given by Theorem 4.3.3. It is thus not surprising that $\widehat{\gamma}_n$ is asymptotically normal with an asymptotic covariance at two different values p and q given by V in (4.14). This result is stated in the following theorem.

Theorem 4.5.1. *Suppose that Assumption 4.3.1 and Assumption 4.5.1 hold. Let $\widehat{\gamma}_n$ denote the empirical autocovariance function of the sample $X_{1:n}$. Then, as $n \to \infty$,*

$$\sqrt{n}\ (\widehat{\gamma}_n - \gamma) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V)\ , \tag{4.29}$$

where V is defined by (4.15). As a consequence, we also have, as $n \to \infty$,

$$\sqrt{n}\ (\widehat{\rho}_n - \rho) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, W)\ , \tag{4.30}$$

where $\widehat{\rho}_n(h) = \widehat{\gamma}_n(h)/\widehat{\gamma}_n(0)$, $\rho(h) = \gamma(h)/\gamma(0)$ and

$$W(p,q) = \sum_{u=1}^{\infty} \{\rho(u+p) + \rho(u-p) - 2\rho(u)\rho(p)\} \times \{\rho(u+q) + \rho(u-q) - 2\rho(u)\rho(q)\}\ . \tag{4.31}$$

Proof. The CLT (4.30) follows from (4.29) (see Exercise 4.3), so we only show (4.29).

As in the proof of Theorem 4.3.3, we can take $\mu = 0$ without loss of generality. We also observe that (4.18) implies that $\widehat{\gamma}_n = \tilde{\gamma}_n + O_p(n^{-1})$. Hence it is sufficient to prove that, as $n \to \infty$,

$$\sqrt{n}\ (\tilde{\gamma}_n - \gamma) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V)\ , \tag{4.32}$$

The proof of this is left to the reader, see Exercise 4.4. □

The asymptotic covariances of the empirical autocovariances V are a bit intricate and depend on η. Surprisingly (at least at first sight) η no longer appears in the asymptotic covariance when one considers the *empirical autocorrelation function* defined by

$$\widehat{\rho}_n(h) = \frac{\widehat{\gamma}_n(h)}{\widehat{\gamma}_n(0)}$$

which is used as an estimator of $\rho(h) = \gamma(h)/\gamma(0)$, where ρ is called the *autocorrelation function*.

4.6 Application to ARMA processes

Let us give some applications of the above asymptotic results on some examples of ARMA processes.

Example 4.6.1 (Strong white noise)**.** *If $(X_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0, \sigma^2)$, we are in the i.i.d. case. Of course Theorem 4.4.2 and Theorem 4.5.1 apply. Note that $\rho(h) = 0$ for all $h \neq 0$ and $W(p,q) = \mathbb{1}_{\{p=q\}}$. Hence (4.30) implies that, for any $K \geq 1$, $\sqrt{n}[\widehat{\rho}_n(1), \ldots, \widehat{\rho}_n(K)]$ converges weakly to an i.i.d. standard Gaussian vector. As a consequence the statistic*

$$T_n = \sum_{l=1}^{K} \widehat{\rho}_n(l)^2$$

converges weakly to a χ^2 ("chi squared") distribution with K degrees of freedom, see Jacod and Protter [2003]. This result can be used to obtain a test of the null hypothesis H_0: "X is a white noise" for a given asymptotic false detection probability*.*

Example 4.6.2 (MA(1) process)**.** *Define X by the non-centered MA(1) equation*

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

where $Z \sim \mathrm{IID}(0, \sigma^2)$. Then the conditions of Theorem 4.4.2 and Theorem 4.5.1 are satisfied. We have $2\pi f(0) = \sigma^2(1+\theta)^2$ and

$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \dfrac{\theta_1}{1+\theta_1^2} & \text{if } |h| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$W(h,h) = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1) & \text{if } |h| = 1 \\ 1 + 2\rho(1)^2 & \text{if } |h| \geq 2 \end{cases}$$

One easily deduces confidence intervals for μ and $\rho(h)$ for given coverage probabilities.

Example 4.6.3 (Empirical mean of an AR(1) process)**.** *Let X be the unique stationary solution of the non-centered AR(1) equation*

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

where $Z \sim \mathrm{IID}(0, \sigma^2)$ and $|\phi| < 1$. Then X has mean μ and autocovariance function given by

$$\gamma(k) = \frac{\sigma^2}{(1-\phi^2)} \phi^{|k|}$$

and its spectral density function reads

$$f(\lambda) = \frac{\sigma^2}{2\pi \left|1 - \phi \mathrm{e}^{-\mathrm{i}\lambda}\right|^2} \ .$$

Then the assumptions of Theorem 4.4.2 are satisfied and the limit variance in (4.23) reads $2\pi f(0) = \sigma^2/(1-\phi)^2$. As a consequence, the confidence interval with asymptotic coverage probability 95% for the mean μ is given by $[\widehat{\mu}_n - 1.96\sigma n^{-1/2}/(1-\phi), \widehat{\mu}_n + 1.96\sigma n^{-1/2}/(1-\phi)]$, hence has maximal size when $\phi \to 1$ and minimal size when $\phi \to -1$.

The assumptions of Theorem 4.5.1 also hold. A direct computation yields

$$\begin{aligned} W(h,h) &= \sum_{m=1}^{h} \phi^{2h}(\phi^{-m} - \phi^m)^2 + \sum_{m=h+1}^{\infty} \phi^{2m}(\phi^{-i} - \phi^i)^2 \\ &= (1-\phi^{2h})(1+\phi^2)(1-\phi^2)^{-1} - 2h\phi^{2h} \end{aligned}$$

4.7 Maximum likelihood estimation

Maximum likelihood estimation is a general approach for the estimation of the parameter in the framework of a dominated model. Let us consider an observed data set, for instance a sample of the $\mathbb{R}^p$-valued time series $(\mathbf{Z}_t)_{t\in\mathbb{Z}}$ at time instants $t = 1, \ldots, n$. We will denote $\mathbf{Z}_{1:n} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. A dominated model means that $\mathbf{Z}_{1:n}$ admits a density $p(\cdot|\theta^*)$ with respect to a known dominating measure, where θ^* is unknown in a given (finite-dimensional) parameter set Θ.

Definition 4.7.1 (Maximum likelihood estimator)**.** *The likelihood of an observation set is defined as the (random) function*

$$\theta \mapsto L_n(\theta) = p(\mathbf{Z}_{1:n}|\theta) \ .$$

The maximum likelihood estimator is then defined as

$$\widehat{\theta}_n \stackrel{\text{def}}{=} \operatorname*{argmin}_{\theta\in\Theta} - \log L_n(\theta) \ , \tag{4.33}$$

when this argmin *is well defined.*

In practice, $\widehat{\theta}_n$ is often obtained through a numerical procedure which, in the best cases, insure that

$$- \log L_n(\widehat{\theta}_n) \leq \inf_{\theta\in\Theta} -L_n(\theta) + o_P(n^{-1/2}) \ . \tag{4.34}$$

To apply such a numerical procedure, the primary question is that of numerically computing the negated log-likelihood $-\log L_n(\theta)$ efficiently for all $\theta \in \Theta$, and for certain procedures its gradient and perhaps also its Hessian.

If $\mathbf{Z}_{1:n}$ is a Gaussian vector, $\mathbf{Z}_{1:n} \sim \mathcal{N}(\boldsymbol{\mu}(\theta^*), \Sigma(\theta^*))$ with $\Sigma(\theta)$ invertible for all $\theta \in \Theta$, then the dominating measure can be taken to be the Lebesgue measure on $\mathbb{R}^n$ and we have

$$- \log L_n(\theta) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det\Sigma(\theta) + \frac{1}{2}\left((\mathbf{Z}_{1:n} - \mu(\theta))^T\Sigma(\theta)^{-1}(\mathbf{Z}_{1:n} - \mu(\theta))\right) \ .$$

This expression is fine if the inverse and the determinant of $\Sigma(t)$ are easily computed, which can become quite a difficult problem if n is large. An alternative, which moreover applies in a more general context than the Gaussian one, is to use the successive conditional density,

$$p(\mathbf{Z}_n|\mathbf{Z}_{1:n-1}, \theta) = \frac{p(\mathbf{Z}_{1:n}|\theta)}{p(\mathbf{Z}_{1:n-1}|\theta)} \ ,$$

with the convention that $p(\mathbf{Z}_1|\mathbf{Z}_{1:0}, \theta) = p(\mathbf{Z}_1|\theta)$. As a function of $\mathbf{Z}_n$, this is a well defined density a.s. in $\mathbf{Z}_{1:n-1}$ and it is the density of the conditional distribution of $\mathbf{Z}_n$ given $\mathbf{Z}_{1:n-1}$ under the parameter θ. It follows that

$$- \log L_n(\theta) = -\sum_{t=1}^{n} \log p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}, \theta) \ . \tag{4.35}$$

Under the *usual regular assumption* (see Anderson [2003]), the Information matrix is defined as

$$I_n(\theta) = \operatorname{Cov}\left(\partial \log L_n(\theta)|\theta\right) = -\mathbb{E}\left[\partial\partial^T \log L_n(\theta)|\theta\right] \ , \tag{4.36}$$

where the mention of θ in the conditional expectation and in the covariance indicate that these are calculated under the distribution given by the parameter θ. As a consequence of (4.35) and (4.36), $I_n(\theta)$ may also be computed as a sum of more elementary terms.

In *nice* models such as i.i.d. regular models (but not only these ones!), the maximum likelihood estimator defined by (4.33) (or satisfying (4.34)) is consistent and asymptoticly normal. Moreover, the information matrix is asymptotically equivalent to $n\mathcal{I}(\theta)$ with $\mathcal{I}(\theta)$ invertible, which provides the asymptotic covariance matrix of the maximum likelihood estimator,

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \Longrightarrow \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta)\right) \tag{4.37}$$

Here we state these facts without details; however, let us stress that such asymptotic results may be quite involved to prove in the dependent case (that is, when $\mathbf{Z}_{1:n}$ is not an i.i.d. sample) or may even fail to hold.

Now, returning to the Gaussian assumption, by Proposition Proposition B.4.8, the conditional density $p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}, \theta)$ is that of $\mathcal{N}\left(\mathbf{Z}_t - \boldsymbol{\eta}_t(\theta), \widetilde{\Sigma}_t(\theta)\right)$ where

$$\boldsymbol{\eta}_t(\theta) = \mathbf{Z}_t - \mathbb{E}\left[\mathbf{Z}_t \middle| \mathbf{Z}_{1:t-1}, \theta\right] , \tag{4.38}$$

$$\widetilde{\Sigma}_t(\theta) = \mathrm{Cov}\left(\mathbf{Z}_t - \boldsymbol{\eta}_t(\theta) | \theta\right) , \tag{4.39}$$

Hence (4.35) yields

$$-2\log L_n(\theta) = n\log(2\pi) + \sum_{t=1}^{n} \log\det\widetilde{\Sigma}_t(\theta) + \sum_{t=1}^{n} \boldsymbol{\eta}_t(\theta)^T \widetilde{\Sigma}_t(\theta)^{-1} \boldsymbol{\eta}_t(\theta) . \tag{4.40}$$

Denoting by ∂_i the derivative with respect to the i-th component of θ, the gradient is then given by

$$-2\partial_i \log L_n = \sum_{t=1}^{n} \left\{ \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\right) + 2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i \boldsymbol{\eta}_t] - \boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \right\} , \tag{4.41}$$

and the Hessian matrix by

$$\begin{aligned} -2\partial_i\partial_j \log L_n = \sum_{t=1}^{n} \Big\{ & \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \widetilde{\Sigma}_t]\right) - \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j \widetilde{\Sigma}_t]\right) \\ & + 2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \boldsymbol{\eta}_t] + 2[\partial_i \boldsymbol{\eta}_t^T]\widetilde{\Sigma}_t^{-1}[\partial_j \boldsymbol{\eta}_t] - 2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j \boldsymbol{\eta}_t] \\ & - 2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_j \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_i \boldsymbol{\eta}_t] - \boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \\ & + 2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \Big\} . \end{aligned} \tag{4.42}$$

Observe that by (4.38), the derivatives of any order of $\boldsymbol{\eta}_t$ with respect to θ are $\sigma(\mathbf{Z}_{1:t-1})$-measurable (the term $\mathbf{Z}_t$ vanishes). On the other hand, $\widetilde{\Sigma}_t$ is deterministic and we have $\mathbb{E}\left[\boldsymbol{\eta}_t(\theta) \middle| \mathbf{Z}_{1:t-1}, \theta\right] = 0$. Hence, when applying this conditional expectation to the summand in (4.42), the following terms vanishes :

$$\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \boldsymbol{\eta}_t], \quad -2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j \boldsymbol{\eta}_t], \quad -2\boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_j \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_i \boldsymbol{\eta}_t] .$$

Now using that $\mathbb{E}[\{\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T\}(\theta)|\theta] = \widetilde{\Sigma}_t(\theta)$, we further have

$$\mathbb{E}\left[\left\{ \boldsymbol{\eta}_t^T \widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \right\}(\theta) \middle| \theta \right] = \mathrm{Trace}\left(\left\{ \widetilde{\Sigma}_t^{-1}[\partial_i\partial_j \widetilde{\Sigma}_t] \right\}(\theta) \right) ,$$

and

$$\mathbb{E}\left[\left\{\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t\right\}(\theta)\middle|\,\theta\right] = \mathrm{Trace}\left(\left\{\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\right\}(\theta)\right) .$$

Using these facts to compute the expectation of $-2\partial_i\partial_j \log L_n$, we finally obtain

$$I_n(i,j;\theta) = \sum_{t=1}^{n}\left\{\mathbb{E}\left[\left\{[\partial_i\boldsymbol{\eta}_t^T]\widetilde{\Sigma}_t^{-1}[\partial_j\boldsymbol{\eta}_t]\right\}(\theta)\middle|\,\theta\right] + \frac{1}{2}\mathrm{Trace}\left(\left\{\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\right\}(\theta)\right)\right\} . \quad (4.43)$$

As a result a meaningful estimator of $I_n(\theta^*)$ is obtained with

$$\widehat{I}_n(i,j) = \sum_{t=1}^{n}\left[\left\{[\partial_i\boldsymbol{\eta}_t^T]\widetilde{\Sigma}_t^{-1}[\partial_j\boldsymbol{\eta}_t]\right\}(\widehat{\theta}_n) + \frac{1}{2}\mathrm{Trace}\left(\left\{\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\right\}(\widehat{\theta}_n)\right)\right] ,$$

and, in view of (4.37), one may use the following approximation to build confidence regions for θ^*,

$$\mathbb{P}(\sqrt{\widehat{I}_n}(\widehat{\theta}_n - \theta^*) \in R) \approx \mathbb{P}(\mathbf{U} \in R) ,$$

where $\mathbf{U} \sim \mathcal{N}(0, I)$ and $\sqrt{\widehat{I}_n}$ is such that $\sqrt{\widehat{I}_n}(\widehat{I}_n)^{-1}\sqrt{\widehat{I}_n}^T$ is the identity matrix (for instance using a Choleski decomposition of $\widehat{I}_n$).

4.8 EM algorithm

In its simpler conditional form (4.35), and even in the Gaussian case (4.40), the log likelihood is often efficiently computed. However, it can be difficult to maximize with a gradient descent algorithm. Such a procedure is iterative and without strong convexity assumptions, each step of a gradient descent does not guaranty an increase of the log likelihood. The EM algorithm, introduced in Dempster et al. [1977], does have this nice property. We here briefly describe this algorithm because it can be quite interesting and practical in the context of state space models.

The EM (*Expectation-Minimization*) algorithm applies in the context of *hidden variables* or *partly observed data*. Consider the model and notation used in Section 4.7.

In this section, we suppose that we can introduce additional variables $\mathbf{U}_n$ for which the joint likelihood $p(\mathbf{Z}_{1:n}, \mathbf{U}_n|\theta)$ is well defined (the density is defined with respect to a dominating measure that do not depend on θ).

If this joint likelihood is easier to compute than the marginal likelihood $p(\mathbf{Z}_{1:n}|\theta)$, the only (!) problem is that since $\mathbf{U}_n$ is not observed, the complete joint likelihood seems to be useless for estimating θ. Nevertheless, given a parameter θ', one can compute

$$\mathcal{Q}_n(;\theta\,;\theta') \stackrel{\text{def}}{=} -\mathbb{E}\left[\log p(\mathbf{Z}_{1:n}, \mathbf{U}_n|\theta)|\,\mathbf{Z}_{1:n}, \theta'\right] , \quad (4.44)$$

from the data, as a function of θ. The EM algorithm is then very simply given as follows.

Algorithm 5: EM algorithm.

Data: Data $\mathbf{Z}_{1:n}$, initial estimate θ_0.
Result: Numerical approximation of the MLE $\widehat{\theta}_n$.
Initialization: Set $k = 0$.
repeat
  Set

$$\theta_{k+1} = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{Q}_n(;\theta\,;\theta_k)\ . \tag{4.45}$$

  Increment k ($k \leftarrow k+1$).
until $\log L_n(\theta_k)$ *and* $\log L_n(\theta_{k-1})$ *are "relatively close"*;

One has the following result.

Theorem 4.8.1. *Define* $(\theta_k)_{k\in\mathbb{N}}$ *iteratively by (4.45). Then* $(L_n(\theta_k))_{k\in\mathbb{N}}$ *is a non-decreasing sequence.*

Proof. We have, for all $\theta, \theta' \in \Theta$,

$$\begin{aligned}
\mathcal{Q}_n(;\theta\,;\theta') - \mathcal{Q}_n(;\theta'\,;\theta') &= \mathbb{E}\left[\log \frac{p(\mathbf{Z}_{1:n}, \mathbf{U}_n|\theta')}{p(\mathbf{Z}_{1:n}, \mathbf{U}_n|\theta)} \,\middle|\, \mathbf{Z}_{1:n}, \theta'\right] \\
&= \log \frac{p(\mathbf{Z}_{1:n}|\theta')}{p(\mathbf{Z}_{1:n}|\theta)} + \mathbb{E}\left[\log \frac{p(\mathbf{U}_n|\mathbf{Z}_{1:n}, \theta')}{p(\mathbf{U}_n|\mathbf{Z}_{1:n}, \theta)} \,\middle|\, \mathbf{Z}_{1:n}, \theta'\right] \\
&\geq \log L_n(\theta') - \log L_n(\theta)\ ,
\end{aligned}$$

since the last term is a Kullback Leibler divergence, which is always nonnegative. It follows that, if θ is chosen so that $\mathcal{Q}_n(;\theta\,;\theta') \leq \mathcal{Q}_n(;\theta'\,;\theta')$, then $L_n(\theta') \leq L_n(\theta)$. Hence the result. □

Of course, in practice, the EM algorithm is used if both the right-hand sides of (4.44) (the *Expectation step*) and (4.45) (the *Minimization step*) are easy to compute numerically.

4.9 Exercises

Exercise 4.1. Let $(\mathbf{X}_n)_{n\in\mathbb{N}}$ and $(\mathbf{Y}_n)_{n\in\mathbb{N}}$ be two sequences of random variables valued in $\mathbb{R}^p$. Denote $\mathbf{Z}_n = \mathbf{X}_n + \mathbf{Y}_n$.

1. Show that $\mathbf{X}_n \xrightarrow{P} 0$ and $\mathbf{Y}_n \xrightarrow{P} 0$ implies $\mathbf{Z}_n \xrightarrow{P} 0$.
2. Show that if $(\mathbf{X}_n)_{n\in\mathbb{N}}$ and $(\mathbf{Y}_n)_{n\in\mathbb{N}}$ are stochastically bounded, then so is $(\mathbf{Z}_n)_{n\in\mathbb{N}}$.
3. In the case where $p = 1$, show that if $X_n \xrightarrow{P} 0$ and $(Y_n)_{n\in\mathbb{N}}$ is stochastically bounded, then $X_n\, Y_n \xrightarrow{P} 0$.

Exercise 4.2. Let $(X_t)_{t\in\mathbb{Z}}$ satisfy Assumption 4.3.1 and Assumption 4.3.2 with $\mu = 0$.

1. Show that (4.11) holds for all $k, l, p, q \in \mathbb{Z}$.

Define

$$A = \sum_{k,\ell,p,q=1}^{m} \sum_{i=-\infty}^{\infty} \psi_{k+i}\psi_{\ell+i}\psi_{p+i}\psi_{q+i}$$

$$B = \sum_{k,\ell,p,q=1}^{m} \{\gamma(k-\ell)\gamma(p-q) + \gamma(k-p)\gamma(\ell-q) + \gamma(k-q)\gamma(\ell-p)\} \ .$$

2. Show that
$$|A| \leq \sum_{k=1}^{m}\sum_{i=\infty}^{\infty} |\psi_{k+i}| \left(\sum_{j=-\infty}^{\infty} |\psi_j|\right)^3 .$$
3. Show that
$$B \leq 3m^2 \left(\sum_{h=-\infty}^{\infty} |\gamma(h)|\right)^2 .$$
4. Conclude that (4.12) holds.

Exercise 4.3. Use the δ-method to show that (4.29) implies (4.30).

Exercise 4.4. Suppose that Assumption 4.3.1 and Assumption 4.5.1 hold with $\mu = 0$. Let $\widehat{\gamma}_n$ denote the empirical autocovariance function of the sample $X_{1:n}$. Let $\tilde{\gamma}_n$ be defined by (4.17). For any $m \geq 1$, define $X^{(m)} = \mathrm{F}_{\psi^m}(Z)$ with ψ^m defined by (4.24) and $\tilde{\gamma}_n^{(m)}$ be defined as in (4.17) with X replaced by $X^{(m)}$.

1. Compute the autocovariance function γ_m of $\left(X_t^{(m)}\right)_{t\in\mathbb{Z}}$.

Let us define, for all $p, q \in \mathbb{Z}$,

$$V_m(p,q) = (\eta-3)\gamma_m(p)\gamma_m(q) + \sum_{u\in\mathbb{Z}} \left[\gamma_m(u)\gamma_m(u-p+q) + \gamma_m(u+q)\gamma_m(u-p)\right] .$$

2. Use Theorem 4.4.3 to show that
$$\sqrt{n}\ \left(\tilde{\gamma}_n^{(m)} - \gamma_m\right) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V_m) \ ,$$
where V_m is defined by (4.15).

3. Proceed as in the proof of Theorem 4.3.3 to show that

$$\lim_{m\to\infty} \limsup_{n\to\infty} n \ \mathrm{Var}\left(\tilde{\gamma}_n - \tilde{\gamma}_n^{(m)}\right) = 0 \ .$$

4. Use Lemma C.1.9 to conclude the proof of Theorem 4.5.1.

Exercise 4.5. Let (X_t) be a weakly stationary real valued process with mean μ and autocovariance function γ. We observe $X_1, \dots, X_n$.

1. Determine the linear unbiased estimator $\hat{\mu}_n$ of μ that minimizes the risk

$$\mathrm{EQM} = \mathbb{E}[(\mu - \hat{\mu}_n)^2].$$

2. Give the corresponding risk.

Exercise 4.6 (AR estimation using moments). Let $(X_t)_{t\in\mathbb{Z}}$ be a real valued centered weakly stationary process with covariance function γ. Denote, for all $t \geq 1$,

$$\Gamma_t = \mathrm{Cov}\left(\left[X_1, \dots, X_t\right]^T\right) = \left[\gamma(i-j)\right]_{1\leq i,j\leq t} \ .$$

Similarly, we define, for all $t \geq 1$,

$$\hat{\Gamma}_t = \left[\hat{\gamma}_n(i-j)\right]_{1\leq i,j\leq t} \ ,$$

where $\hat{\gamma}_n$ is the empirical autocovariance function of the sample $X_1, \dots, X_n$.

1. Show that empirical covariance matrices $\hat{\Gamma}_t$ are invertible for all $t \geq 1$ under a simple condition on $X_1, \dots, X_n$. [Hint : use that $\hat{\gamma}_n$ is a nonnegative definite hermitian function and Exercise 2.9.]

Consider the AR(p) process

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \varepsilon_t$$

where $(\varepsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$. Suppose that we observe a sample $X_1, \dots, X_n$ of this process.

2. Define a *moment estimator* of $\phi_1, \dots, \phi_p$ and σ^2 by solving the Yule-Walker equations with γ replaced by the empirical autocovariance function $\hat{\gamma}_n$. Show that this approach does provide uniquely defined estimators $\hat{\phi}_1, \dots, \hat{\phi}_p$ and $\hat{\sigma}^2$.

3. Show that the operator $\hat{\Phi}(\mathrm{B}) = 1 - \sum_{k=1}^{p} \hat{\phi}_k \, \mathrm{B}^k$ is causally invertible.

4. Give a condition on $\phi_1, \dots, \phi_p$ for which this method appears to be appropriate.

Exercise 4.7 (Likelihood of Gaussian processes). Consider n observations $X_1, \dots, X_n$ from a regular, centered, 2nd order stationary Gaussian process with autocovariance function γ_θ depending on an unknown parameter $\theta \in \Theta$. For an assumed value of θ, define the following innovation sequence

$$\begin{cases} I_{1,\theta} = X_1, \quad v_{1,\theta} = \gamma_\theta(0) \\ I_{t,\theta} = X_t - \hat{X}_{t,\theta}, \quad v_{t,\theta} = \mathrm{Var}(I_{t,\theta}|\theta) \quad \text{for } t = 2, \dots, n \end{cases}$$

where $\hat{X}_{t,\theta}$ denotes the L^2 projection of X_t onto $\mathrm{Span}(X_1, \dots, X_{t-1})$ and $\mathrm{Var}(\cdot|\theta)$ the variance, under the distribution of parameter θ.

1. Show that the log-likelihood of θ can be written as

$$\log p(X_1, \dots, X_n|\theta) = -\frac{1}{2}\left[n\log(2\pi) + \sum_{t=1}^{n}\left\{\log v_{t,\theta} + \frac{I_{t,\theta}^2}{v_{t,\theta}}\right\}\right]$$

2. Consider the AR(1) model $X_t = \phi X_{t-1} + \varepsilon_t$ where (ε_t) is a Gaussian white noise of variance σ^2 and define $\theta = (\phi, \sigma^2)$ and $\Theta = (-1,1) \times (0,\infty)$. Show that the log-likelihood then satisfies

$$\log p_\theta(X_1, \dots, X_n) = -\frac{1}{2}\Bigg[n\log(2\pi) + \log\left(\frac{\sigma^2}{1-\phi^2}\right) + \frac{X_1^2(1-\phi^2)}{\sigma^2} + (n-1)\log\sigma^2 + \sum_{t=2}^{n}\frac{(X_t - \phi X_{t-1})^2}{\sigma^2}\Bigg]$$

   Deduce the expression of the "conditional" maximum likelihood estimator $\hat{\theta}_n = (\hat{\phi}_n, \hat{\sigma}_n^2)$, obtained by maximizing $\log p_\theta(X_2, \dots, X_n|X_1)$.

3. How to handle the case where $\Theta = [-1,1]^c \times (0,\infty)$ or, more generally, $\Theta = (\mathbb{R} \setminus \{-1,1\}) \times (0,\infty)$?

In the following, we assume that $\Theta = (-1,1) \times (0,\infty)$.

4. Show that the Fisher information matrix for θ is equivalent to nJ when $n \to \infty$, where J is a matrix to be determined.

We admit that the maximum likelihood estimator is asymptotically efficient, that is,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_2(0, J^{-1}) \ .$$

5. Construct an asymptotic test for testing the null hypothesis $H_0 : \phi = 0$ against the alternative $H_1 : \phi \neq 0$ at asymptotic level $\alpha \in (0,1)$. That is, find a statistic T_n and a decision threshold t_α such that, under the null hypothesis,

$$\lim_{n\to\infty} \mathbb{P}(T_n > t_\alpha) = \alpha \ .$$

   The decision threshold t_α will be expressed as a quantile of the $\mathcal{N}(0,1)$ law.

We now consider the MA(1) model $X_t = \varepsilon_t + \rho\varepsilon_{t-1}$, where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a centered Gaussian white noise with variance σ^2 and $\theta = (\rho, \sigma^2) \in \Theta = (-1,1) \times (0,\infty)$.

6. Show that the innovation sequence can be computed according to the following recursion:

$$\begin{cases} I_{1,\theta} = X_1, \quad v_{1,\theta} = (1+\rho^2)\sigma^2 \\ I_{t,\theta} = X_t - \rho\frac{\sigma^2}{v_{t-1,\theta}} I_{t-1,\theta}, \quad v_{t,\theta} = (1+\rho^2)\sigma^2 - \frac{\rho^2\sigma^4}{v_{t-1,\theta}} \quad \text{for } t = 2, \dots, n \end{cases}$$

7. Considering $\tilde{v}_{t,\theta} = v_{t,\theta}/\sigma^2$, obtain the expression of $\hat{\sigma}_n^2$ as a function of $\hat{\rho}_n$ and of the observations $X_1, \dots, X_n$.

8. Show that, for all $\theta \in \Theta$, $\tilde{v}_{t,\theta} \to 1$.

Part II

Financial time series

Chapter 5

Stochastic autoregressive models

So far we mainly discussed models generated in a linear way (such as ARMA processes) and studied these models almost exclusively through their second order properties (mean and variance). The Gaussian assumption is very convenient in this context. The Gaussian distribution is stable through linear transformation, it is uniquely defined using second order properties and conditional distributions can be computed based on linear projection. It provides a large class of processes which admit a linear representation. Indeed, if a Gaussian time series $X = (X_t)_{t\in\mathbb{Z}}$ is regular and purely non-deterministic, then its wold decomposition in Section 2.6 gives that

$$X_t = \sum_{s\in\mathbb{Z}} \psi_s Z_{t-s} \ ,$$

where Z is a Gaussian white noise, hence a strong white noise.

Unfortunately, in many practical situations, linear models and the Gaussian assumption are not adapted to the structure of the data. The goal of this chapter is to provide an introduction to alternative approaches. Of course, we will not pretend to be exhaustive.

5.1 Standard models for financial time series

5.1.1 Conditional volatility

Financial time series such as stock indices or currency exchange rates (see Figure 1.5) are not correctly modeled by stationary processes. In fact, it is more meaningful to model the *returns* of a financial time series $(p_t)_{t\in\mathbb{Z}}$. Here t generally refers to time indices of daily measurements (then p_t is the closing or opening price of working day t). Higher sampling frequency can be considered, but, below one hour, different models are necessary to take into account trading mechanisms and their impact on short range price jumps. The returns are defined as the relative increments

$$s_t = (p_t - p_{t-1})/p_{t-1} \ ,$$

which provides a scale free measure of the evolution of the price p_t between time $t-1$ and t. In practice it is more convenient to use the *log-returns* which are simply defined as

$$r_t = \log p_t - \log p_{t-1} \ , \quad t \in \mathbb{Z} \ .$$

Note that the two definitions are related, since we have

$$r_t = \log(1 + (p_t - p_{t-1})/p_{t-1}) = \ln(1 + s_t) = s_t(1 + o(1)) \ ,$$

as r_t or $s_t \to 0$, which corresponds to the usual situation where s_t is much smaller than one in magnitude. In fact the log returns r_t can be seen as a continuous time equivalent of the returns s_t. To see why, consider the same relative wealth that would be obtained from time $t-1$ to time t by *compounding* m steps with constant returns equal to $m^{-1}s_t^{(m)}$, that is

$$1 + s_t = (1 + m^{-1}s_t^{(m)})^m = \mathrm{e}^{s_t^{(m)}(1+o(1))} \quad \text{as } m \to \infty \,.$$

Taking the logarithm on both sides we see that r_t is the limit of $s_t^{(m)}$ as $m \to \infty$. This why the log returns are also sometimes referred to as *continuous compounding* returns.

In econometrics, the *volatility* of a price generally refers to a measure of the local variations of its returns or log-returns, which is often seen as a rough measurement of the risk related to a given asset. The *conditional volatility* is more precisely defined as follows.

Definition 5.1.1 (Conditional volatility)**.** *Let $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $(r_t)_{t\in\mathbb{Z}}$ be a time series adapted to this filtration. The* conditional volatility *of the sequence $(r_t)_{t\in\mathbb{Z}}$ with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is defined as the square-rooted sequence of conditional variances $\left(\sqrt{\mathrm{Var}\,(r_t \,|\, \mathcal{F}_{t-1})}\right)_{t\in\mathbb{Z}}$.*

If $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is the natural filtration $\mathcal{F}_t = \sigma(r_s,\, s \le t)$ of $(r_t)_{t\in\mathbb{Z}}$ the considered conditional volatility is purely *endogenous*. With this definition of the filtration, typical financial time series are observed to have a very small conditional mean $\mathbb{E}\,[r_t|\,\mathcal{F}_{t-1}]$. Supposing that it vanishes for all t, in which case one says that the time series is a series of *martingale differences*. Observe that a weakly stationary time series of martingale differences is a weak white noise. As a consequence ARMA models as defined in Section 3.3 appears to be useless to model such time series. Similarly, if a stationary Gaussian process is a series of martingale differences, it also boils down to a centered strong white noise. Thus standard time series models such as ARMA or Gaussian processes are not well suited for modelling financial time series viewed as martingale increments. It also indicates that such models have poor interests for forecasting the returns, linearly or not, from the past. Nevertheless, these models can be developed for forecasting the conditional volatility as we shall see hereafter.

5.1.2 ARCH and GARCH processes

An *autoregressive conditionally heteroscedastic* (ARCH) process is probably the simpler approach to provide a closed form formula to the endogenous conditional volatility.

Definition 5.1.2 (ARCH process)**.** *Let $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. A process $(Y_t)_{t\in\mathbb{Z}}$ which is adapted to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is said to be an ARCH(p) process if there exists $a > 0$ and coefficients $b_1, \ldots, b_p \ge 0$ such that, for all $t \in \mathbb{Z}$,*

$$\mathbb{E}\,[Y_t|\,\mathcal{F}_{t-1}] = 0 \,. \tag{5.1}$$

$$\mathbb{E}\left[Y_t^2\middle|\,\mathcal{F}_{t-1}\right] = a + \sum_{k=1}^{p} b_k Y_{t-k}^2 \,. \tag{5.2}$$

For ARCH processes, the conditional volatility is necessarily endogenous since if (5.1) and (5.2) hold with $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ with respect to which $(Y_t)_{t\in\mathbb{Z}}$ is adapted, then they also hold with the natural filtration.

Observe that Relation (5.2) guarantees that the conditional variance is positive, whatever the past of Y is, and is expressed in a very simple way. The fact that it depends only on a finite number of past observations can be seen as a too important restriction. This is the main reason why *generalized* autoregressive conditionally heteroscedastic (GARCH) processes were introduced. They extend ARCH processes by only adding a finite number of parameters.

Definition 5.1.3 (GARCH process)**.** *Let $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. A process $(Y_t)_{t\in\mathbb{Z}}$ which is adapted to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is said to be a GARCH(p, q) process if there exist $a > 0$ and coefficients $b_1, \ldots, b_p, c_1, \ldots, c_q \geq 0$ such that (5.1) holds and*

$$\mathbb{E}\left[Y_t^2 \middle| \mathcal{F}_{t-1}\right] = \sigma_t^2 \quad \textit{with} \quad \sigma_t^2 = a + \sum_{k=1}^{p} b_k Y_{t-k}^2 + \sum_{k=1}^{q} c_k \sigma_{t-k}^2 \,. \tag{5.3}$$

Taking the expectation in (5.1) and (5.3), we get that a weakly stationary GARCH process $(Y_t)_{t\in\mathbb{Z}}$ has zero mean and variance σ^2 given by

$$\sigma^2 = \frac{a}{1 - \sum_{k=1}^{p} b_k - \sum_{k=1}^{q} c_k} \,, \tag{5.4}$$

implying that

$$\sum_{k=1}^{p} b_k + \sum_{k=1}^{q} c_k < 1 \,. \tag{5.5}$$

In contrast with ARMA processes, because the involved equations (5.2) and (5.3) are non linear, it is not so easy to generate ARCH and GARCH processes from a given white noise. This will be done in Section 5.4.

However it is interesting to mention that ARCH and GARCH processes are intimately related to AR and ARMA processes though the following result.

Proposition 5.1.1. *Let $(Y_t)_{t\in\mathbb{Z}}$ be a weakly stationary GARCH process satisfying (5.1) and (5.3) for some $a > 0$ and $b_1, \ldots, b_p, c_1, \ldots, c_q \geq 0$. Let us assume that $p = q$.Let σ^2 be defined by (5.4). Define, for all $t \in \mathbb{Z}$, $\epsilon_t = Y_t^2 - \sigma_t^2$ and $U_t = Y_t^2 - \sigma^2$. Then $(\epsilon_t)_{t\in\mathbb{Z}}$ is a sequence of martingale increments, and $(U_t)_{t\in\mathbb{Z}}$ is centered and satisfies the following ARMA equation for all $t \in \mathbb{Z}$,*

$$U_t = \sum_{k=1}^{p} (b_k + c_k) U_{t-k} + \epsilon_t - \sum_{k=1}^{p} c_k \, \epsilon_{t-k} \,. \tag{5.6}$$

It is an AR equation if $(Y_t)_{t\in\mathbb{Z}}$ is an ARCH process, that is, $c_1 = \cdots = c_q = 0$.

Proof. For all $t \in \mathbb{Z}$, we have, since $\sigma_t = \mathbb{E}\left[Y_t^2 \middle| \mathcal{F}_{t-1}\right]$ is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}\left[\epsilon_t \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[Y_t^2 \middle| \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\sigma_t^2 \middle| \mathcal{F}_{t-1}\right] = 0 \,.$$

Thus $(\epsilon_t)_{t\in\mathbb{Z}}$, which is adapted to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$, is a sequence of martingale increments. Now, using (5.3) and the definition of ϵ_t, we have

$$\begin{aligned} Y_t^2 - \epsilon_t &= a + \sum_{k=1}^{p} b_k Y_{t-k}^2 + \sum_{k=1}^{q} c_k (Y_{t-k}^2 - \epsilon_{t-k}) \\ &= a + \sum_{k=1}^{p} (b_k + c_k) Y_{t-k}^2 - \sum_{k=1}^{q} c_k \epsilon_{t-k} \,. \end{aligned}$$

Using the definition of U_t, we get

$$U_t + \sigma^2 = a + \sum_{k=1}^{p}(b_k + c_k)(U_{t-k} + \sigma^2) + \epsilon_t - \sum_{k=1}^{q} c_k \epsilon_{t-k} \ .$$

Finally (5.4) yields (5.6). □

Remark 5.1.1. *The assumption $p = q$ in Proposition 5.1.1 can be made without loss of generality by completing the coefficients b_ks or c_ks with zeros up to $k = p \vee q$.*

Remark 5.1.2. *We show in Proposition 5.1.1 that $(\epsilon_t)_{t\in\mathbb{Z}}$ is a martingale increment sequence. To get a (centered) white noise, one needs also to have a constant variance for this sequence. This is of course the case if it is stationary and L^2, see Section 5.4 for practical conditions to get such solutions.*

5.1.3 Stochastic Volatility Models

A *stochastic volatility* (SV) process is defined as follows.

Definition 5.1.4 (Stochastic volatility processes)**.** *A process $(X_t)_{t\in\mathbb{Z}}$ is said to be a stochastic volatility process if there exists two independent processes $(Z_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ such that $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$ and*

$$X_t = \mathrm{e}^{Y_t/2} Z_t \ . \tag{5.7}$$

Setting $\mathcal{F}_t = \sigma(Y_{s+1}, Z_s, s \le t)$ we have that $(X_t)_{t\in\mathbb{Z}}$ is adapted to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ and, for all $t \in \mathbb{Z}$,

$$\begin{aligned} \mathbb{E}\left[X_t \middle| \mathcal{F}_{t-1}\right] &= 0 \ , \\ \mathbb{E}\left[X_t^2 \middle| \mathcal{F}_{t-1}\right] &= \exp(Y_t) \ . \end{aligned}$$

For this reason, the SV model is often presented using the conditional volatility $\sigma_t = \exp(Y_t/2)$.

A stationary SV model is easier to define than stationary ARCH or GARCH processes. It suffices to have a stationary process $\left(\begin{bmatrix} Y_t & Z_t \end{bmatrix}^T\right)_{t\in\mathbb{Z}}$ and then to apply (5.7). We shall see in Section 5.4 that the standard constructions of GARCH processes are such that the conditional volatility may be defined using the natural filtration of $(X_t)_{t\in\mathbb{Z}}$, that is, we can express σ_t^2 as a deterministic function of the past of X up to time $t-1$. This is no longer the case for the SV model, where $\sigma_t = \exp(Y_t/2)$ generally includes some innovation which makes it non-measurable with respect to $\mathcal{F}_{t-1} \supset \sigma(X_s, s \le t-1)$.

The SV models bear some similarities with the state space approach developed in Assumption 7.1.1. As in Definition 7.1.1, the observed time series $(X_t)_{t\in\mathbb{Z}}$ is defined using an observation equation (5.7) with a white noise $(Z_t)_{t\in\mathbb{Z}}$ and *state* (one also says *latent* or *hidden*) variables $(Y_t)_{t\in\mathbb{Z}}$, which follows its own model. A common approach is to model $(Y_t)_{t\in\mathbb{Z}}$ using use a similar state equation as in DLMs, that is an AR model. Taking an AR(1), the complete model is then defined by the state and observation equations

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + W_t \ , \tag{5.8}$$

$$\log(X_t^2) = Y_t + \log(Z_t^2) \ , \tag{5.9}$$

where $(Z_t)_{t\in\mathbb{Z}} \sim \text{IID}(0,1)$ and $(W_t)_{t\in\mathbb{Z}} \sim \text{IID}(0,\sigma^2)$ are two independent white noise. Here we expressed the observation equation by squaring and then taking the log of (5.7), so that it becomes a linear equation for the observation $\log(X_t^2)$. However, because of the logarithmic form of the additive noise in the observation equation, the Gaussian assumption, which is usual in a DLM approach, is generally excluded in this particular application on SV models.

In Figure 5.1 we compare the paths of the absolute log-returns of a financial time series with those obtained by simulating a GARCH(1,1) model and an SV model of the form (5.8)–(5.9). The parameters of the model have been chosen to fit the financial time series (we do not detail how, although this would be an interesting point to develop). The paths are simulated independently and thus are not expected to be close to each other. Here we wish to visualize some common stylized facts on these paths. An important one is that the high values of the series appear in *clusters*. An interpretation of this is to say that periods of high volatility and low volatility alternate along the time, depending on the economic context. One can observe that this phenomenon seems to be amplified for the GARCH series.

Simulating an SV model given by (5.8)–(5.9) is an easy task. We will see in Section 5.4 how to simulate the path of a GARCH process.

5.2 Stochastic autoregressive models

We now introduce a class of processes which extends the AR processes studied in Section 3.3.3.

Definition 5.2.1 (Stochastic autoregressive models)**.** *Let $p \geq 1$ and $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}^T\right)_{t\in\mathbb{Z}}$ be an i.i.d. sequence valued in $\mathbb{R}^p \times \mathbb{R}^{p\times p}$. A time series $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ valued in $\mathbb{R}^p$ is said to be a stochastic autoregressive time series with random matrices $(\mathbf{\Phi}_t)_{t\in\mathbb{Z}}$ and noise sequence $(\mathbf{W}_t)_{t\in\mathbb{Z}}$ if, for all $t \in \mathbb{Z}$, it satisfies the* stochastic autoregressive equation

$$\mathbf{X}_t = \mathbf{\Phi}_t \mathbf{X}_{t-1} + \mathbf{W}_t \,, \tag{5.10}$$

We will say that a solution $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ of (5.10) is non-anticipative *if for all $t \in \mathbb{Z}$, $\mathbf{X}_t$ is $\sigma(\mathbf{W}_s, \mathbf{\Phi}_s,\, s \leq t)$–measurable.*

We will often consider a stationary non-anticipative solution of this equation. In this case, it is tempting to see $(\mathbf{W}_t)_{t\in\mathbb{Z}}$ as the *innovation process* of $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ as in the AR case. However the situation is more complicated here because $\mathbf{\Phi}_t$ will in general be correlated with $\mathbf{W}_t$.

In contrast with the state equation (7.1) of a DLM (or the AR equation (3.20)), the matrices $\mathbf{\Phi}_t$ are random and may depend on the innovation $\mathbf{W}_t$. In addition one usually assumes that the innovation $\mathbf{W}_t$ is Gaussian in a DLM while in a random coefficient autoregressive time series, this assumption is much less natural as it no more implies $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ to be Gaussian.

As for the ARMA processes, given an initial condition $\mathbf{X}_s = \mathbf{x}$ at some time instant s and a sequence $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}^T\right)_{t\in\mathbb{Z}}$, there is a unique solution to (5.10) for $t > s$. Indeed, iterating (5.10) for $t = s+1, s+2, \ldots$, this solution is given for all $t > s$ by

$$\mathbf{X}_t(\mathbf{x}) = \mathbf{W}_t + \sum_{j=0}^{t-s-2} \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \ldots \mathbf{\Phi}_{t-j} \mathbf{W}_{t-j-1} + \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \ldots \mathbf{\Phi}_{s+1} \mathbf{x} \,. \tag{5.11}$$

Now, as for the ARMA processes, we shall study the existence and uniqueness of stationary solutions of the random autoregression equation (5.10). However, in contrast with our study

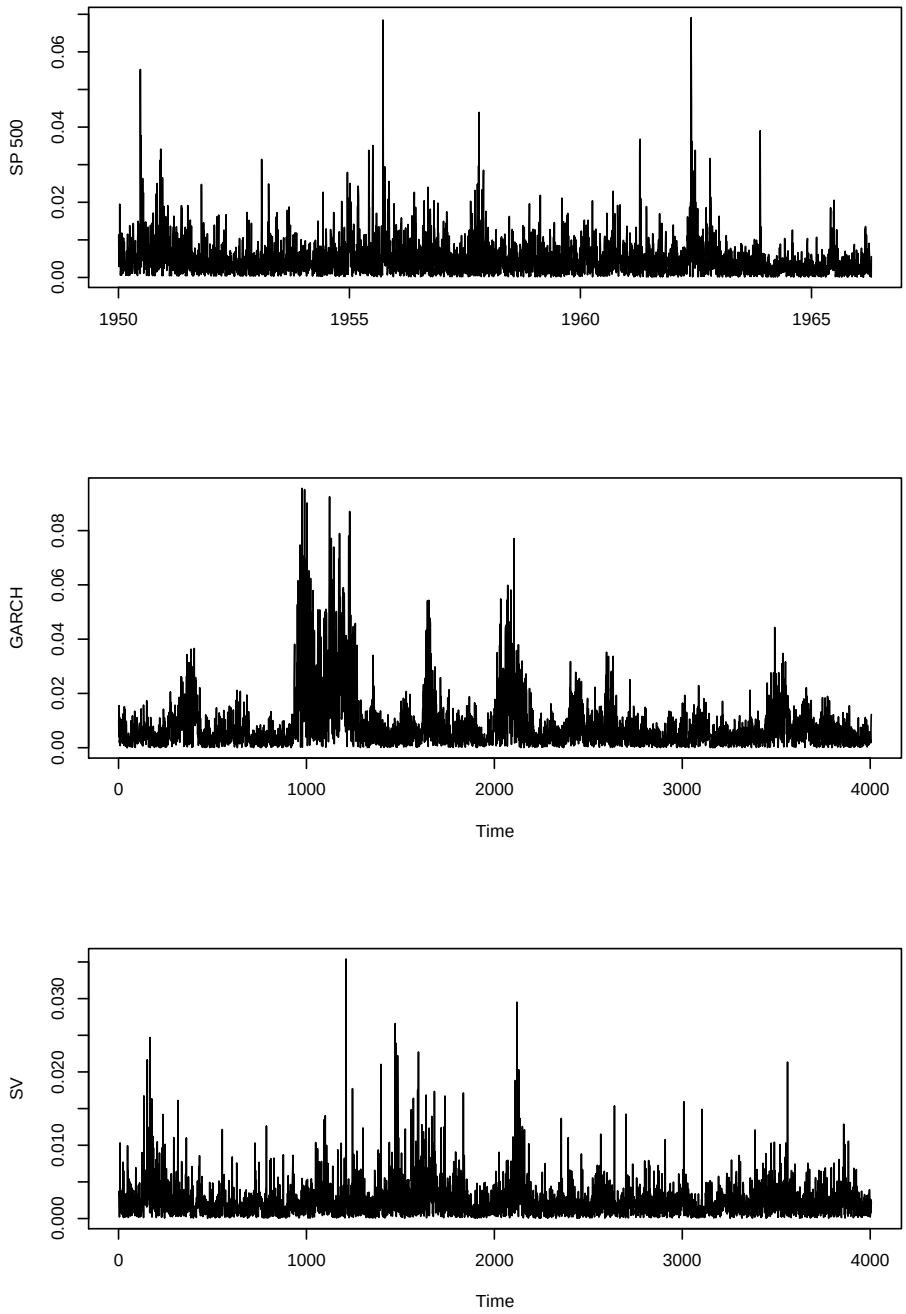

Figure 5.1: Top : Absolute log-returns of the Standard & Poor 500 index from 1950 to 1966. Middle : absolute values of a simulated GARCH(1,1) process with parameters fitted using the SP 500 series. Bottom : absolute values of a simulated SV model with parameters fitted using the SP 500 series.

of ARMA process in Section 3.3, we will use assumptions which guaranty these solutions to be non-anticipative, as defined above.

As seen in (5.11), an essential ingredient is the stability of the product $\mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\dots\mathbf{\Phi}_{t-j}$ as $j\to\infty$. Hence the following assumptions, the first one to find a strictly stationary solution and the second one to get an additional control of the L^ℓ norm.

Assumption 5.2.1. *The i.i.d. sequence* $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}^T\right)_{t\in\mathbb{Z}}$ *satisfies*

$$\mathbb{E}\left[\log(|\mathbf{W}_0|\vee 1)\right]<\infty$$
$$\mathbb{E}\left[\log(|\mathbf{\Phi}_0|\vee 1)\right]<\infty \quad \textit{and} \quad \limsup_{n\to\infty} n^{-1}\mathbb{E}\left[\log|\mathbf{\Phi}_1\mathbf{\Phi}_2\dots\mathbf{\Phi}_n|\right]<0\,. \tag{5.12}$$

Here $|\cdot|$ denotes the Euclidean operator norm on $\mathbb{R}^{p\times p}$ matrices, $|A|=\sup_{x\in\mathbb{R}^p}|Ax|$. However, clearly, Assumption 5.2.1 and Assumption 5.2.2 do not depend on the choice of the norm. In fact, because the sequence defined by the log in (5.12) is subadditive ($a_{n+m}\le a_n+a_m$ the lim sup in (5.12) is a limit. Moreover, Kingman's subadditive ergodic Theorem (see Kingman [1973]) (or the law of large numbers if $p=1$) implies the following lemma.

Lemma 5.2.1. *If* $(\mathbf{\Phi}_k)_{k\ge1}$ *is i.i.d. and (5.12) holds, then, there exists* $\alpha>0$ *such that, for all* $t\in\mathbb{Z}$,

$$\limsup_{n\to\infty} n^{-1}\log|\mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\dots\mathbf{\Phi}_{t-n}|\le-\alpha \quad a.s.$$

Proof. Let $t\in\mathbb{Z}$ and set $U_n=\log|\mathbf{\Phi}_t\mathbf{\Phi}_{t+1}\dots\mathbf{\Phi}_{t+n-1}|$. Then using that $|\cdot|$ is a matrix norm, we have, for all $n,m\ge1$,

$$\begin{aligned}U_{n+m}&\le\log|\mathbf{\Phi}_t\mathbf{\Phi}_{t+1}\dots\mathbf{\Phi}_{t+n-1}|+\log|\mathbf{\Phi}_{t+n}\mathbf{\Phi}_{t+1}\dots\mathbf{\Phi}_{t+n+m-1}|\\&\le U_n+U_m\circ\mathrm{S}^n\ .\end{aligned}$$

Here U_n is seen as function of $(\mathbf{\Phi}_s)_{s\in\mathbb{Z}}$ and S is the usual shift operator. The sequence $(U)_{n\ge1}$ satisfying the inequality $U_{n+m}\le U_n+U_m\circ\mathrm{S}^n$ is called subadditive. The left-hand part of (5.12) moreover says that $\mathbb{E}\left[U_1\vee0\right]<\infty$. By Kingman's subadditive ergodic Theorem (see Kingman [1973]), we get that

$$\lim_{n\to\infty} n^{-1}U_n=\inf_{n\ge1} n^{-1}\mathbb{E}\left[U_n\right]\quad\text{a.s.}\,.$$

Using that the (nonrandom) sequence $(u_n)_{n\ge1}$ with $u_n=\mathbb{E}\left[U_n\right]$ is subbadditive it is a standard exercise to show that $n^{-1}u_n$ converges to $\inf_{n\ge1}n^{-1}u_n$ as $n\to\infty$, hence

$$\lim_{n\to\infty} n^{-1}U_n=\lim_{n\to\infty} n^{-1}u_n=:\alpha\quad\text{a.s.}\,.$$

Now, observe that

$$u_n=\mathbb{E}\left[U_n\right]=\mathbb{E}\left[U_n\circ\mathrm{S}^{1-t}\right]=\mathbb{E}\left[\log|\mathbf{\Phi}_1\mathbf{\Phi}_2\dots\mathbf{\Phi}_n|\right]\ .$$

Hence, by left-hand part of (5.12) we have $\alpha<0$. We conclude the proof by observing that $(\mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\dots\mathbf{\Phi}_{t-n})_{n\ge1}$ has the same distribution as $(\mathbf{\Phi}_t\mathbf{\Phi}_{t+1}\dots\mathbf{\Phi}_{t+n})_{n\ge1}$. □

To obtain some control on the moments of the solution we shall consider a similar assumption involving moments. We use the following definition.

Assumption 5.2.2. *Let $\ell \geq 1$. The i.i.d. sequence $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}^T\right)_{t\in\mathbb{Z}}$ satisfies*

$$\mathbb{E}\left[|\mathbf{W}_0|^\ell\right] < \infty \,,$$
$$\mathbb{E}\left[|\mathbf{\Phi}_0|^\ell\right] < \infty \quad \textit{and} \quad \limsup_{n\to\infty} n^{-1} \log \mathbb{E}\left[|\mathbf{\Phi}_1\mathbf{\Phi}_2\ldots\mathbf{\Phi}_n|^\ell\right] < 0 \,. \tag{5.13}$$

By Jensen's Inequality, we see that Assumption 5.2.2 implies Assumption 5.2.1. We make this assumption in the following for sake of simplicity.

We can now state the following result.

Theorem 5.2.2. *Let $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}\right)_{t\in\mathbb{Z}}$ be an i.i.d. sequence valued in $\mathbb{R}^p \times \mathbb{R}^{p\times p}$ satisfying Assumption 5.2.1. Then there exists a unique solution of (5.10) which is bounded in probability at $-\infty$ ($\mathbf{X}_t = O_P(1)$ as $t \to -\infty$). Moreover this solution is non-anticipative and strictly stationary, hence it is the unique stationary solution. If moreover Assumption 5.2.2 holds for some $\ell \geq 1$, then all the complex eigenvalues of $\mathbb{E}[\mathbf{\Phi}_0]$ have modulus strictly less than 1 and the unique stationary solution is uniformly bounded in L^ℓ norm and has (finite) mean*

$$\boldsymbol{\mu} = (I - \mathbb{E}[\mathbf{\Phi}_0])^{-1}\,\mathbb{E}[\mathbf{W}_0] \,. \tag{5.14}$$

Proof. If we can show that for any given $t \in \mathbb{Z}$ the sum

$$\sum_{j=0}^{\infty} \mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\ldots\mathbf{\Phi}_{t-j}\mathbf{W}_{t-j-1}$$

is absolutely convergent a.s., then is easy to get that defining the random variables

$$\mathbf{X}_t = \mathbf{W}_t + \sum_{j=0}^{\infty} \mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\ldots\mathbf{\Phi}_{t-j}\mathbf{W}_{t-j-1} \,, \tag{5.15}$$

the process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is strictly stationary and is a non-anticipative solution to Equation (5.10) for all $t \in \mathbb{Z}$.

Note that, by Lemma 5.2.1, Condition (5.12) implies that there exists some (random) constant $C > 0$ such that

$$|\mathbf{\Phi}_t\mathbf{\Phi}_{t-1}\ldots\mathbf{\Phi}_{t-j}| \leq C\,\mathrm{e}^{-\alpha j} \quad \text{a.s.} \tag{5.16}$$

Denote $\mathbf{w}_k = 1 \vee |\mathbf{W}_k|$, which is a random variable larger than or equal to 1 and to $|\mathbf{W}_k|$. Observe that

$$\begin{aligned}\sum_{j\geq 0} \mathbb{P}(\mathrm{e}^{-\alpha j/2}\mathbf{w}_{t-j-1} > 1) &\leq \mathbb{E}\left[\sum_{j\geq 0} \mathbb{1}\{\mathbf{w}_{t-j-1} > \mathrm{e}^{\alpha j/2}\}\right] \\ &\leq \mathbb{E}\left[\sum_{j\geq 0} \mathbb{1}\{j \leq 2\log(\mathbf{w}_{t-j-1})/\alpha\}\right] \\ &\leq 1 + 2\mathbb{E}\left[\log(\mathbf{w}_{t-j-1})\right]/\alpha \,,\end{aligned}$$

which is finite by Assumption 5.2.1. Hence, by Borel-Cantelli's Lemma, we get that, for all $t \in \mathbb{Z}$, almost surely, for j large enough, $\mathbf{w}_{t-j-1} < \mathrm{e}^{\alpha j/2}$. With (5.16) we get that the

series $\sum_{j=0}^{\infty} \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \dots \mathbf{\Phi}_{t-j} \mathbf{W}_{t-j-1}$ is absolutely convergent a.s., which concludes the proof for showing the existence of astrictly stationary non-anticipative solution.

Let us now prove the uniqueness. Let $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$ be any solution of (5.10) which is uniformly bounded in probability at $-\infty$. Applying (5.11) to $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ and $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$, we obtain that, for all $s < t \in \mathbb{Z}$,

$$\mathbf{X}_t - \mathbf{Y}_t = \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \dots \mathbf{\Phi}_{s+1} (\mathbf{X}_s - \mathbf{Y}_s) \; . \tag{5.17}$$

Using (5.16) and that $\mathbf{X}_s$ and $\mathbf{Y}_s$ are bounded in probability as $s \to -\infty$, we get that $\mathbf{X}_t - \mathbf{Y}_t = o_P(1)$ as $s \to -\infty$. Thus $\mathbf{X}_t = \mathbf{Y}_t$ a.s.

Suppose now that Assumption 5.2.2 holds. Then we have , for all $t \in \mathbb{Z}$ and $j \geq 0$,

$$\begin{aligned} \mathbb{E}\left[\left| \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \dots \mathbf{\Phi}_{t-j} \mathbf{W}_{t-j-1} \right|^{\ell} \right] &\leq \mathbb{E}\left[\left| \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \dots \mathbf{\Phi}_{t-j} \right|^{\ell} \right] \mathbb{E}\left[\left| \mathbf{W}_{t-j-1} \right|^{\ell} \right] \\ &\leq C \, \mathrm{e}^{-\gamma n} \; , \end{aligned} \tag{5.18}$$

for some positive constants C and γ depending neither on t nor on j. It follows that, for all $t \in \mathbb{Z}$, the series $\sum_{j=0}^{\infty} \mathbf{\Phi}_t \mathbf{\Phi}_{t-1} \dots \mathbf{\Phi}_{t-j} \mathbf{W}_{t-j-1}$ is absolutely convergent in $L^{\ell}(\Omega, \mathcal{F}, \mathbb{P})$, and thus (5.15) holds in the L^{ℓ} sense. Moreover we have, using (5.18),

$$\mathbb{E}\left[\left| \mathbf{X}_t \right|^{\ell} \right]^{1/\ell} \leq \mathbb{E}\left[\left| \mathbf{W}_t^{\ell} \right] \right]^{1/\ell} + \sum_{j=0}^{\infty} C \, \mathrm{e}^{-\gamma n/\ell} \; .$$

Observe that the right-hand side is finitely bounded independently of t. Thus $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is uniformly bounded in L^{ℓ} norm. Since $\ell \geq 1$, the series also converges in L^1 norm and we can take the expectation in the right-hand side of (5.15) and obtain

$$\boldsymbol{\mu} = \mathbb{E}\left[\mathbf{W}_0 \right] + \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbf{\Phi}_n \mathbf{\Phi}_{n-1} \dots \mathbf{\Phi}_1 \mathbf{W}_0 \right] \; .$$

Using that the sequence $\left(\begin{bmatrix} \mathbf{W}_t & \mathbf{\Phi}_t \end{bmatrix}^T \right)_{t\in\mathbb{Z}}$ is i.i.d., we get

$$\boldsymbol{\mu} = \left(I + \sum_{n=1}^{\infty} \left(\mathbb{E}\left[\mathbf{\Phi}_0 \right] \right)^n \right) \mathbb{E}\left[\mathbf{W}_0 \right] \; .$$

Moreover the sum between parentheses, say S, is an absolutely convergent series of matrices and satisfies $S(I - \mathbb{E}\left[\mathbf{\Phi}_0 \right]) = I$. Hence we get (5.14). □

In the next sections, we examine some classical applications.

5.3 Examples

5.3.1 AR processes

We mentioned that stochastic autoregressive models extend the AR processes introduced in Section 3.3. Recall the AR(p) equation

$$Y_t = \phi_1 Y_{t-1} + \dots \phi_p Y_{t-p} + Z_t \; , \tag{5.19}$$

where the innovation $(Z_t)_{t\in\mathbb{Z}}$ is a strong white noise. Such an equation can indeed be formulated as a vector AR equation (thus a particular case of (5.10)) by introducing the lag-vector

$$\mathbf{X}_t = \begin{bmatrix} Y_t & Y_{t-1} & \dots & Y_{t-p+1} \end{bmatrix}^T , \tag{5.20}$$

so that (5.19) can be rewritten as

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{W}_t , \tag{5.21}$$

where $\mathbf{W}_t = \begin{bmatrix} Z_t & 0 & \dots & 0 \end{bmatrix}^T$ and Φ is the deterministic companion matrix

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \dots & \phi_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix} .$$

Denoting $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ the associated AR polynomial, we note that the eigenvalues of Φ are the reciprocals of the zeros of ϕ. Observe that Condition (5.12) applies for (5.21) if and only if the largest modulus of all its eigenvalues is in $[0, 1)$, which, in turns, is equivalent to have that the polynomial ϕ does not vanish on the unit disk.

5.3.2 Stochastic autoregressive equations of order p

The example of Section 5.3.1 suggests a way to express a stochastic autoregressive equation or order $p \geq 2$ as a stochastic autoregressive equation of order 1 in a similar way. A process $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$ will be said to follow a stochastic autoregressive equation or order $p \geq 2$ if, for all $t \in \mathbb{Z}$,

$$\mathbf{Y}_t = \sum_{i=1}^{p} \mathbf{\Psi}_{i,t} \mathbf{Y}_{t-i} + \mathbf{Z}_t , \tag{5.22}$$

where $\left(\begin{bmatrix} \mathbf{Z}_t & \mathbf{\Psi}_{1,t} & \dots & \mathbf{\Psi}_{p,t} \end{bmatrix}_t^T \right)_{t\in\mathbb{Z}}$ is an i.i.d. sequence valued in $\mathbb{R}^q \times \mathbb{R}^{q\times q\times p}$. Again, we may rewrite this equation as a stochastic autoregressive equations of order 1 (5.10) by setting $\mathbf{W}_t = \begin{bmatrix} \mathbf{Z}_t^T & 0 & \dots & 0 \end{bmatrix}^T$ and using the block companion matrix

$$\mathbf{\Phi}_t = \begin{bmatrix} \mathbf{\Psi}_{1,t} & \mathbf{\Psi}_{2,t} & \mathbf{\Psi}_{3,t} & \dots & \mathbf{\Psi}_{p,t} \\ \mathbf{1} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{1} & 0 \end{bmatrix} ,$$

where $\mathbf{1}$ denotes the (here $q \times q$) identity matrix.

5.3.3 Bilinear processes

A bilinear process $(Y_t)_{t\in\mathbb{Z}}$ is defined as the solution of a bilinear equation of the form

$$Y_t = \sum_{j=1}^{p} a_j Y_{t-j} + Z_t + \sum_{j=1}^{p} b_j Y_{t-j} Z_{t-1} \tag{5.23}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is a strong white-noise. Then, using the lag-vector defined as in (5.20), the process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is a solution of a stochastic autoregressive equation (5.10) with $\mathbf{W}_t = \begin{bmatrix} Z_t & 0 & \dots & 0 \end{bmatrix}^T$ and

$$\mathbf{\Phi}_t = \begin{bmatrix} \phi_{1,t} & \phi_{2,t} & \cdots & \phi_{p,t} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \tag{5.24}$$

where $\phi_{j,t} = a_j + b_j Z_{t-1}$.

5.3.4 Discrete autoregressive processes

There are several approaches to model discrete data with some autoregressive structure. Let us first mention the INteger AutoRegressive processes (INAR processes) , although they cannot be expressed as the solutions of a stochastic autoregressive equation of the form (5.10). Consider an array of i.i.d. Bernoulli random variables $(U_{t,k})_{t\in\mathbb{Z},k\geq 1}$ with mean θ and an integer valued i.i.d. process $(Z_t)_{t\in\mathbb{Z}}$, independent of $(U_{t,k})_{t\in\mathbb{Z},k\geq 1}$. An INAR process is defined as the solution of the stochastic equation

$$X_t = \sum_{k=1}^{X_{t-1}} U_{t,k} + Z_t \tag{5.25}$$

Such process are studied for instance in Al-Osh and Alzaid [1987]. Here we consider a simpler class of processes, the *Discrete AutoRegressive* processes.

Definition 5.3.1 (DAR processes)**.** *Let p be a positive integer and $\left(\begin{bmatrix} Z_t & U_t & \ell_t \end{bmatrix}\right)_{n\in\mathbb{N}}$ be an i.i.d. sequence valued in $\mathbb{N}\times\{0,1\}\times\{1,\dots,p\}$. A DAR(p) process $(X_t)_{t\in\mathbb{Z}}$ is a solution to the equation*

$$Y_t = U_t Y_{t-\ell_t} + (1-U_t)Z_t\ . \tag{5.26}$$

We can again express Equation (5.26) as a stochastic autoregressive equation (5.10) using the lag-vector (5.20). Then it suffices to set, for all $t\in\mathbb{Z}$, $\mathbf{W}_t = \begin{bmatrix} (1-U_t)Z_t & 0 & \dots 0 \end{bmatrix}^T$ and

$$\mathbf{\Phi}_t = \begin{bmatrix} \phi_{1,t} & \phi_{2,t} & \cdots & \phi_{p,t} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \end{bmatrix},$$

where $\phi_{j,t} = U_t \mathbb{1}_{\{j\}}(\ell_t)$ for $j = 1,\dots,p$.

5.4 Application to GARCH processes

5.4.1 GARCH processes generated by stochastic recursions

We shall use the results of Section 5.2 to define stationary ARCH and GARCH processes. We first explain how stochastic recursions yields ARCH and GARCH processes.

Proposition 5.4.1. *Let $a > 0$, $b_1, \dots, b_p \geq 0$ and $(\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$. Let $(Y_t)_{t\in\mathbb{Z}}$ be an L^2 non-anticipative solution of the stochastic recursion*

$$Y_t = \sigma_t \epsilon_t \tag{5.27}$$

$$\sigma_t^2 = a + \sum_{k=1}^{p} b_k Y_{t-k}^2 \,, \tag{5.28}$$

where σ_t is taken positive. Here non-anticipative *means that $(Y_t)_{t\in\mathbb{Z}}$ is adapted to the natural filtration of $(\epsilon_t)_{t\in\mathbb{Z}}$, Then $(Y_t)_{t\in\mathbb{Z}}$ is an ARCH process with coefficients $a, b_1, \dots, b_p$, that is, it satisfies (5.1) and (5.2) with $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ defined as the natural filtration of $(\epsilon_t)_{t\in\mathbb{Z}}$.*

For ARCH processes, Equation (5.28) is simple in the sense that $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ only appear on the left-hand side of the equation. Defining $(Y_t)_{t\in\mathbb{Z}}$ as a solution of (5.27) and (5.28) does not raise confusion. Indeed Equations (5.27) and (5.28) can equivalently be written as

$$Y_t = \left(a + \sum_{k=1}^{p} b_k Y_{t-k}^2 \right)^{1/2} \epsilon_t \,.$$

It seems more convenient, however, to express theses equations in terms of $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ rather than $(Y_t)_{t\in\mathbb{Z}}$, that is, as

$$\sigma_t^2 = a + \sum_{k=1}^{p} b_k \sigma_{t-k}^2 \epsilon_{t-k}^2 \,. \tag{5.29}$$

Note, however, that $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{Z}}$ instead of $(\mathcal{F}_t)_{t\in\mathbb{Z}}$. The GARCH equation is more complicated. To avoid confusions in this case, it is better to define $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ alone as a solution of a stochastic recurrent equation and then $(Y_t)_{t\in\mathbb{Z}}$ by (5.27).

Proposition 5.4.2. *Let $a > 0$, $b_1, \dots, b_p \geq 0$ and $(\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$. Let $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ be an L^1 non-anticipative solution of the stochastic recursion*

$$\sigma_t^2 = a + \sum_{k=1}^{p} b_k \sigma_{t-k}^2 \epsilon_{t-k}^2 + \sum_{k=1}^{q} c_k \sigma_{t-k}^2 \,. \tag{5.30}$$

Here non-anticipative *means that $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{Z}}$, where $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is the natural filtration of $(\epsilon_t)_{t\in\mathbb{Z}}$. Define $(Y_t)_{t\in\mathbb{Z}}$ by (5.27) with σ_t positive. Then $(Y_t)_{t\in\mathbb{Z}}$ is a GARCH process with coefficients $a, b_1, \dots, b_p, c_1, \dots, c_q$, that is, it satisfies (5.1) and (5.3) with $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ defined as above.*

The proofs of Proposition 5.4.1 and Proposition 5.4.2 are left to the reader (Exercise 5.1). These results open the way to the construction of stationary ARCH and GARCH processes. We start with ARCH processes and then consider GARCH(1,1) and finally GARCH(p,q) processes.

5.4.2 GARCH$(1,1)$ case

We start with the case of GARCH$(1,1)$ processes. They are the simplest GARCH processes in the sense that they can be written as the solutions of a scalar stochastic autoregressive

equation. Indeed the GARCH(1,1) stochastic recursion expressed on the volatility as in (5.30) reads as

$$\sigma_t^2 = a + (b\epsilon_{t-1}^2 + c)\sigma_{t-1}^2 \,, \tag{5.31}$$

for some $a > 0$ and $b, c \geq 0$. This is exactly (5.10) with $\mathbf{X}_t = \sigma_t^2$, $\mathbf{W}_t = a$ and $\mathbf{\Phi}_t = (b\epsilon_{t-1}^2 + c)$. In this case, by the law of large numbers (see Theorem 4.1.5) Assumption 5.2.1 holds if and only if

$$\mathbb{E}\left[\log(b\epsilon_0^2 + c)\right] < 0 \,, \tag{5.32}$$

and Assumption 5.2.2 holds with $\ell = 1$ if and only if

$$\mathbb{E}\left[b\epsilon_0^2 + c\right] < 1 \,,$$

which is a stronger assumption by Jensen's Inequality, as expected. When $(\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$ the later condition reads $b + c < 1$. Recall that we already mentioned in (5.5) that $b + c < 1$ is a necessary condition on the GARCH(1,1) coefficients in Definition 5.1.3. Applying Theorem 5.2.2, we obtain the following result.

Theorem 5.4.3. *Let $a > 0$ and $b, c \geq 0$. Suppose that $(\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$ and that (5.32) holds. Then there exists a unique solution $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ of (5.31) which is uniformly bounded in probability at $-\infty$. The resulting process $(Y_t)_{t\in\mathbb{Z}}$ defined by (5.27) is the unique stationary non-anticipative solution to (5.27) and (5.31) and it is L^2 if and only if $b + c < 1$.*

5.4.3 General case

We now extend Theorem 5.4.3 to GARCH(p, q) processes. This case is more involved as we need to rely on a more complicated stochastic autoregression representation. In this context we rewrite the GARCH equations (5.27) and (5.30) as a stochastic autoregressive equation (5.10) by setting

$$\mathbf{X}_t = \begin{bmatrix} Y_t^2 & Y_{t-1}^2 & \dots & Y_{t-p+1}^2 & \sigma_t^2 & \sigma_{t-1}^2 & \dots & \sigma_{t-q+1}^2 \end{bmatrix}^T \tag{5.33}$$

$$\mathbf{W}_t = \begin{bmatrix} a\epsilon_t^2 & 0 & \dots & 0 & a & 0 & \dots & 0 \end{bmatrix} \tag{5.34}$$

$$\mathbf{\Phi}_t = \begin{bmatrix} \begin{bmatrix} b_1\epsilon_t^2 & \dots & \dots & \dots & b_p\epsilon_t^2 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} c_1\epsilon_t^2 & & \dots & & c_q\epsilon_t^2 \\ 0 & & \dots & & 0 \\ \vdots & & & & \vdots \\ 0 & & \dots & & 0 \end{bmatrix} \\ \begin{bmatrix} b_1 & & \cdots & & b_p \\ 0 & & \dots & & 0 \\ \vdots & & & & \vdots \\ 0 & & \dots & & 0 \end{bmatrix} & \begin{bmatrix} c_1 & \dots & \dots & \dots & c_q \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \end{bmatrix} . \tag{5.35}$$

We obtain the following result by applying Theorem 5.2.2.

Theorem 5.4.4. *Let $a > 0$, $p, q \geq 1$ and $b_1, \dots, b_p \geq 0$, $c_1, \dots, c_q \geq 0$. Suppose that $(\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,1)$. Then there is a non-anticipative weakly stationary solution $(Y_t)_{t\in\mathbb{Z}}$ to (5.27) and (5.30) if and only if $b_1 + \cdots + b_p + c_1 + \cdots + c_q < 1$. Moreover, if this condition holds, it is strictly stationary and it is the unique solution for which $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is uniformly bounded in probability at $-\infty$.*

Proof. The existence of a GARCH process with coefficients $b_1, \dots, b_p \geq 0$, $c_1, \dots, c_q \geq 0$ implies $b_1 + \cdots + b_p + c_1 + \cdots + c_q < 1$, see (5.5). Hence the necessary condition.

We now suppose that $b_1 + \cdots + b_p + c_1 + \cdots + c_q < 1$, that is, Condition (5.5) holds. To any solution $(Y_t)_{t\in\mathbb{Z}}$ of (5.27) and (5.30) corresponds a solution $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ of the stochastic autoregressive equation (5.10) through the definitions (5.33)–(5.35). The result is then a consequence of Theorem 5.2.2, provided that we are able to show that Assumption 5.2.2 holds with $\ell = 1$ (which is stronger that Assumption 5.2.1). We have $\mathbb{E}\left[|\mathbf{W}_0|\right] < \infty$ so it only remains to show (5.13) with $\ell = 1$, and since we are in the i.i.d. case we can forget the $\sup_t$ in (5.13) and set $t = 0$. For a matrix $M = [M_{i,j}]_{1\leq i,j\leq p}$, we define the ℓ^1 norm as

$$|M|_1 = \sum_{i,j=1}^{p} |M_{i,j}| \, .$$

Observe that the matrix $\mathbf{\Phi}_t$ in (5.35) has non-negative entries. Thus for all $n \geq 1$, $\mathbf{\Phi}_n \dots \mathbf{\Phi}_1$ also has non-negative entries. It follows that

$$\mathbb{E}\left[|\mathbf{\Phi}_n \dots \mathbf{\Phi}_1|_1\right] = \left|\mathbb{E}\left[\mathbf{\Phi}_n \dots \mathbf{\Phi}_1\right]\right|_1 = \left|\left(\mathbb{E}\left[\mathbf{\Phi}\right]\right)^n\right|_1 \, ,$$

where $\mathbf{\Phi}$ denotes a generic random matrix with same distribution as the $\mathbf{\Phi}_t$s. Using (5.35), we find

$$\mathbb{E}\left[\mathbf{\Phi}\right] = \begin{bmatrix} \begin{bmatrix} b_1 & \dots & \dots & \dots & b_p \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} c_1 & & \dots & & c_q \\ 0 & & \dots & & 0 \\ \vdots & & & & \vdots \\ 0 & & \dots & & 0 \end{bmatrix} \\ \begin{bmatrix} b_1 & & \cdots & & b_p \\ 0 & & \dots & & 0 \\ \vdots & & & & \vdots \\ 0 & & \dots & & 0 \end{bmatrix} & \begin{bmatrix} c_1 & \dots & \dots & \dots & c_q \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \end{bmatrix} .$$

whose characteristic polynomial reads (up to a positive or negative sign)

$$P(\lambda) = \lambda^{p+q} \left(1 - \sum_{k=1}^{p} b_k \lambda^{-k} - \sum_{k=1}^{q} c_k \lambda^{-k}\right) \, .$$

Using (5.5), we see that $P(\lambda)$ does not vanish for $|\lambda| \geq 1$ and the spectral radius of $\mathbb{E}\left[\mathbf{\Phi}\right]$ is less than one. Thus there exists $\lambda \in (0,1)$ and $C > 0$ such that

$$\left|\left(\mathbb{E}\left[\mathbf{\Phi}\right]\right)^n\right|_1 \leq C \, \lambda^n \, .$$

We thus get that Condition (5.13) holds with $\ell = 1$. This yields Assumption 5.2.2 (and thus also Assumption 5.2.1). Hence Theorem 5.2.2 applies. □

5.5 Uniformly bounded solutions of stochastic autoregressive equations

Some arguments of the proof of Theorem 5.2.2 can be carried on when $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}^T\right)_{t\in\mathbb{Z}}$ is not i.i.d. Instead we can rely on the notion of *uniformly bounded* sequences.

Definition 5.5.1. *A $\mathbb{R}^p$-valued time series $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is said to be uniformly bounded in L^ℓ-norm if*

$$\sup_{t\in\mathbb{Z}} \|\mathbf{X}_t\|_\ell < \infty \ ,$$

where we denote by $\|\cdot\|_\ell$ the norm of the $L^\ell(\Omega,\mathcal{F},\mathbb{P})$ space for $\ell\in[1,\infty]$.

Then Assumption 5.2.2 can be weakened (dropping the i.i.d. assumption) into

Assumption 5.5.1. *Let $\ell,\ell'\in[1,\infty]$. The sequence $(\mathbf{W}_t)_{t\in\mathbb{Z}}$ is uniformly bounded in L^ℓ-norm and $(\mathbf{\Phi}_t)_{t\in\mathbb{Z}}$ is uniformly bounded in $L^{\ell'}$-norm and*

$$\limsup_{n\to\infty} n^{-1} \sup_{s\in\mathbb{Z}} \log \|\mathbf{\Phi}_{s+n}\mathbf{\Phi}_{s+n-1}\dots\mathbf{\Phi}_{s+1}\|_{\ell'} < 0 \ . \tag{5.36}$$

Then Theorem 5.2.2 can be adapted to this set of assumption as follows.

Theorem 5.5.1. *Suppose that $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{\Phi}_t\end{bmatrix}\right)_{t\in\mathbb{Z}}$ satisfies Assumption 5.5.1. Then there exists a unique solution of (5.10) which is bounded in probability at $-\infty$ ($\mathbf{X}_t = O_P(1)$ as $t\to-\infty$). Moreover this solution is non-anticipative and uniformly bounded in L^{ℓ_0} norm, where $\ell_0\in[1,\infty]$ with $1/\ell_0 = 1/\ell + 1/\ell'$, hence is the unique solution which is uniformly bounded in L^{ℓ_0} norm.*

Proof. See Exercise 5.5. □

An immediate application of Theorem 5.5.1 is the case of the AR processes, by relying on the description of Section 5.3.1 but this time allowing $(Z_t)_{t\in\mathbb{Z}}$ to be a weak white noise. In this case, the solution of Theorem 5.5.1 is the unique weakly stationary solution and thus it exactly corresponds to the case of causal AR processes as defined in Section 3.3.

More interestingly, Theorem 5.5.1 applies in the case of a non-stationary AR(1) equation. Consider a sequence $(\phi_t)_{t\in\mathbb{Z}}$ such that

$$\sup_{t\in\mathbb{Z}} |\phi_t| < 1 \ .$$

Let $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0,\sigma^2)$. Then Theorem 5.5.1 implies that there is a unique solution of the equation

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

which is uniformly bounded in L^2 norm.

5.6 Exercises

Exercise 5.1. Prove Proposition 5.4.1 and Proposition 5.4.2, successively.

Exercise 5.2 (A special construction of an ARCH(p) process). Let $p \geq 1$, $a > 0$, and $b_1, b_2, \ldots, b_p \geq 0$. We consider the stationary solution of the following non-centered AR equation

$$Y_t = a + \sum_{i=1}^{p} b_i Y_{t-i} + U_t$$

where $(U_t)_{t\in\mathbb{Z}} \sim \text{IID}(0, \sigma^2)$. We assume that $\Phi(1) > 0$, where $\Phi(z) = 1 - \sum_{k=1}^{p} b_i z^k$.

1. Show that the filter associated to $\Phi(\text{B})$ is causally invertible.

2. Show that $(Y_t)_{t\in\mathbb{Z}}$ is a positively valued process under appropriate conditions on the support of the marginal distribution of $(U_t)_{t\in\mathbb{Z}}$.

Let $(V_t)_{t\in\mathbb{Z}}$ be an i.i.d. centered process such that $\mathbb{P}(V_t = \pm 1) = 1/2$ which is independent of the $(U_t)_{t\in\mathbb{Z}}$. We let $X_t = V_t\sqrt{Y_t}$ for all $t \in \mathbb{Z}$.

3. Show that $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process.

Exercise 5.3 (Kurtosis of the conditionally Gaussian GARCH(p, q) process). Let $p, q \geq 1$, $a > 0$, $b_1, b_2, \ldots, b_p \geq 0$, and $(\epsilon_t)_{t\in\mathbb{Z}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We assume that $b_1 + \cdots + b_p + c_1 + \cdots + c_q < 1$ and $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is the stationary solution of (5.30) and, define $(Y_t)_{t\in\mathbb{Z}}$ by (5.27). Let $\mathcal{F}_t$ denote the σ-field generated by $(\epsilon_s)_{s\leq t}$.

1. Justify that $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is adapted to $(\mathcal{F}_{t-1})_{t\in\mathbb{Z}}$.

2. What is the conditional distribution of Y_t given $\mathcal{F}_{t-1}$?

From now on we assume $\mathbb{E}[\sigma_t^4] < \infty$. Denote by $\kappa = \mathbb{E}[Y_t^4]/(\mathbb{E}[Y_t^2])^2$ the *kurtosis* of Y_t.

3. Show that

$$\kappa = 3 + 3\frac{\text{Var}(\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}])}{(\mathbb{E}[Y_t^2])^2} \ .$$

[Hint : Recall that for a $\mathcal{N}(0, 1)$-distributed random variable Z, we have $\mathbb{E}[Z^4] = 3$.]

4. For a GARCH(1,1) process, with $p = q = 1$ and $a, b = b_1, c = c_1 > 0$, show that

$$\kappa = 3 + \frac{6b^2}{1 - (c^2 + 2bc + 3b^2)} \ .$$

Exercise 5.4 (Even moments of the conditionally Gaussian GARCH(1, 1) process). Let $a > 0$, $b, c \geq 0$ and $(\epsilon_t)_{t\in\mathbb{Z}} \sim \text{IID}(0, 1)$. We assume that Condition (5.32) holds and $\left(\sigma_t^2\right)_{t\in\mathbb{Z}}$ is the stationary solution of (5.31), and define $(Y_t)_{t\in\mathbb{Z}}$ by (5.27).

1. What is the solution if $b = 0$? In the following we assume $b > 0$.

2. Provide an expansion formula for σ_t^2.

3. Show that for all integer $r \geq 1$, $\mathbb{E}[|Y_t|^{2r}] < \infty$ if and only if

$$c^r + \sum_{j=1}^{r} \frac{r!}{j!(r-j)!} \mathbb{E}\left[\epsilon_0^{2j}\right] \; b^j c^{r-j} < 1 \; .$$

4. Precise this condition in the case where $(\epsilon_t)_{t\in\mathbb{Z}}$ is Gaussian.

Exercise 5.5. Prove Theorem 5.5.1. Start by showing a bound similar to (5.18) under Assumption 5.5.1 using the Hölder inequality

$$\|XY\|_{\ell_0} \leq \|X\|_{\ell} \, \|Y\|_{\ell'} \; .$$

Deduce the existence of a solution which is uniformly bounded in L^{ℓ_0} norm. Use (5.17) to obtain the uniqueness.

Chapter 6

Weakly stationary multivariate time series

In this chapter, we consider random variables valued in Hilbert spaces. We can see those as radom processes index by the Hilbert space. We introduce the covariance operator in this context, a special case of which are the usual covariance matrices of random vectors. Next we introduce random process in Hilbert spaces and in particular the weakly stationary time series. This includes the case of *multivariate* time series.

6.1 L^2 random variables valued in a Hilbert space

We extend the definition of L^2 $\mathbb{C}$-valued random variables as follows.

Definition 6.1.1 (Hilbert valued r.v.)**.** *Let $\mathcal{H}_0$ be a Hilbert space and $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. A $\mathcal{H}_0$-valued r.v. defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is a measurable mapping from $(\Omega, \mathcal{F})$ to $\mathcal{H}_0$ endowed with the Borel σ-field associated to the Hilbert norm $|\cdot|$ on $\mathcal{H}_0$. The set of $\mathcal{H}_0$-vaued random variables X such that*

$$|X|_{\mathcal{H}} := \mathbb{E}\left[|X|^2_{\mathcal{H}_0}\right]^{1/2} < \infty$$

is denoted by $L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$. It is a Hilbet space $\mathcal{H}$, once endowed with the inner product

$$\langle X, Y\rangle_{\mathcal{H}} = \mathbb{E}\left[\langle X, Y\rangle_{\mathcal{H}_0}\right] .$$

We will in fact mainly deal with the finite dimensional case.

Example 6.1.1 (Mutlivariate r.v.)**.** Multivariate r.v. *The simplest example, and the main one in these lecture notes is the case where $\mathcal{H}_0$ has finite dimension, say $\mathcal{H}_0 = \mathbb{C}^d$ for some $d \geq 1$. Then a $\mathcal{H}_0$-valued r.v. X simply is a* multivariate *random variable, $X = (X_1, \ldots, X_d)$, which we will identify to the column vector $\begin{bmatrix} X_1 & \ldots & X_d \end{bmatrix}^T$, and, in this case,*

$$\langle X, Y\rangle_{\mathcal{H}} = \mathbb{E}\left[\langle X, Y\rangle_{\mathcal{H}_0}\right] = \mathbb{E}\left[X^T \bar{Y}\right] = \sum_{i=1}^{d} \mathbb{E}\left[X_i \bar{Y_i}\right] .$$

One can also consider infinite dimensional spaces.

Example 6.1.2 (Functional r.v.)**.** Functional r.v. *Take $\mathcal{H}_0 = L^2([0,1])$ (in fact all infinite dimensional separable Hilbert spaces are isomorphic, and this example is often seen as the generic example). Let $(\phi_n)_{n\in\mathbb{N}}$ be a Hilbert basis of $\mathcal{H}_0$ and define*

$$X(t) = \sum_{j\in\mathbb{N}} \langle X, \phi_j\rangle_{\mathcal{H}_0}\, \phi_j(t)\ .$$

Then $(\omega, t) \mapsto X(t,\omega)$ is measurable as a $([0,1]\times\Omega, \mathcal{B}([0,1])\otimes\mathcal{F}) \to (\mathbb{C}, \mathcal{B}(\mathbb{C}))$ function and we have

$$\langle X, Y\rangle_{\mathcal{H}} = \mathbb{E}\left[\langle X, Y\rangle_{\mathcal{H}_0}\right] = \mathbb{E}\left[\int_0^1 X(t)\bar{Y}(t)\,\mathrm{d}t\right] = \int_0^1 \mathbb{E}\left[X(t)\bar{Y}(t)\right]\,\mathrm{d}t\ .$$

One can actually often rely on the Riesz representation theorem in order to manipulate r.v. in $\mathcal{H}$ using complex valued r.v. only. Consider for instance the notion of centering. How do we define $\mathbb{E}[X]$ for $X \in \mathcal{H}$? Clearly in the above examples, it is natural to define $\mathbb{E}[X]$ either as the vector of expectations (for Example 6.1.1) or the function $t \mapsto \mathbb{E}[X(t)]$ (for Example 6.1.2). However it is preferable to have a more generic approach.

Definition 6.1.2 (Expectation of a Hilbert-valued r.v.)**.** *Let $\mathcal{H}_0$ be a Hilbert space. Suppose that X is a r.v. in $\mathcal{H} = L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$. Then*

$$x \mapsto \mathbb{E}\left[\langle x, X\rangle_{\mathcal{H}_0}\right]$$

is a linear continuous form on $\mathcal{H}_0$ and, by the Riesz representation theorem, we define $\mathbb{E}[X]$ as the element of $\mathcal{H}_0$ such that, for all $x \in \mathcal{H}_0$,

$$\mathbb{E}\left[\langle x, X\rangle_{\mathcal{H}_0}\right] = \langle x, \mathbb{E}[X]\rangle_{\mathcal{H}_0}\ .$$

Then $X \mapsto \mathbb{E}[X]$ is linear continuous from $\mathcal{H}$ to $\mathcal{H}_0$ and

$$|\mathbb{E}[X]|_{\mathcal{H}_0} \le \mathbb{E}\left[|X|_{\mathcal{H}_0}\right] \le |X|_{\mathcal{H}}\ .$$

We now extend the definition of the variance of Hilbert valued r.v.

Definition 6.1.3 (Variance of a Hilbert-valued r.v.)**.** *Let $\mathcal{H}_0$ be a Hilbert space. Suppose that X is a r.v. in $\mathcal{H} = L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$. Then we define*

$$\mathrm{Var}\,(X) := \|X - \mathbb{E}[X]\|_{\mathcal{H}}^2 = \|X\|_{\mathcal{H}}^2 - \|\mathbb{E}[X]\|_{\mathcal{H}_0}^2$$

6.2 Covariance operator

We can use the same trick as in Definition 6.1.2, this time in order to define the covariance between two L^2 r.v. valued in a Hilbert spaces. The expectation were defined as an element of $\mathcal{H}_0$, the covariance, in contrast is defined as an operator from $\mathcal{G}_0$ to $\mathcal{H}_0$ acting on the (possibly different) corresponding Hilbert spaces on each side of it.

Definition 6.2.1 (Covariance operator)**.** *Let $\mathcal{H}_0$ and $\mathcal{G}_0$ be two Hilbert spaces. Suppose that X and Y are r.v. in $\mathcal{H} = L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} = L^2(\Omega, \mathcal{G}_0, \mathcal{F}, \mathbb{P})$. Then*

$$(x, y) \mapsto \mathrm{Cov}\left(\langle X, x\rangle_{\mathcal{H}_0}, \langle Y, y\rangle_{\mathcal{H}_0}\right)$$

is a sesqui-linear continuous form on $\mathcal{H}_0 \times \mathcal{G}_0$ *and, by the Riesz representation theorem, we define* $\mathrm{Cov}\,(X,Y)$ *as the continuous linear operator defined on* $\mathcal{G}_0$ *by the identity, valid for all* $x \in \mathcal{H}_0$ *and* $y \in \mathcal{G}_0$*,*

$$\langle \mathrm{Cov}\,(X,Y)\,(y), x\rangle_{\mathcal{H}_0} = \mathrm{Cov}\left(\langle X,x\rangle_{\mathcal{H}_0}, \langle Y,y\rangle_{\mathcal{G}_0}\right) .$$

Or using the notation for any linear operator $\Gamma : \mathcal{G}_0 \to \mathcal{H}_0$*,*

$$x^H\,\Gamma\,y = \langle \Gamma(y), x\rangle_{\mathcal{H}_0} ,$$

where x^H *denotes the linear form* $\mathcal{H}_0 \ni y \mapsto \langle y,x\rangle_{\mathcal{H}_0}$*, we write directly*

$$x^H\,\mathrm{Cov}\,(X,Y)\;y = \mathrm{Cov}\left(\langle X,x\rangle_{\mathcal{H}_0}, \langle Y,y\rangle_{\mathcal{G}_0}\right) .$$

We have that, for any $x \in \mathcal{H}_0$ *and* $y \in \mathcal{G}_0$*,*

$$\mathrm{Cov}\,(X+x, Y+y) = \mathrm{Cov}\,(X,Y) .$$

and

$$\|\mathrm{Cov}\,(X,Y)\| := \sup_{y\in\mathcal{G}_0,\,|y|_{\mathcal{G}_0}\le 1} |\mathrm{Cov}\,(X,Y)\,(y)|_{\mathcal{H}_0} \le \mathbb{E}\left[|X|_{\mathcal{H}_0}|Y|_{\mathcal{G}_0}\right] \le |X|_{\mathcal{H}}\;|Y|_{\mathcal{G}} , \tag{6.1}$$

or, replacing X *by* $X - \mathbb{E}\,[X]$*, and the same for* Y*,*

$$\|\mathrm{Cov}\,(X,Y)\| \le \mathrm{Var}\,(X)\;\mathrm{Var}\,(Y) , \tag{6.2}$$

The mapping $(X,Y) \mapsto \mathrm{Cov}\,(X,Y)$ *is a continuous sesquilinear operator from* $\mathcal{H}\times\mathcal{G}$ *to the space of continuous linear operators* $\mathcal{L}_b(\mathcal{G}_0,\mathcal{H}_0)$ *endowed with the* $\|\cdot\|$ *norm.*

If $X = Y$, we often simply denote

$$\mathrm{Cov}(X) = \mathrm{Cov}\,(X,X) ,$$

and call $\mathrm{Cov}(X)$ the *autocovariance* operator of X, which should not be confused with $\mathrm{Var}\,(X)$, which is a nonnegative number and is the same as $\mathrm{Cov}(X)$ only in the univariate case (in fact one can show that $\mathrm{Var}\,(X)$ is the trace of the operator $\mathrm{Cov}(X)$, see Section 6.4). Moreover in this special case we easily see that the autocovariance operator of X is non-negative definite Hermitian, that is, for all $x, y \in \mathcal{H}_0$,

$$x^H\;\mathrm{Cov}(X)\,y = \overline{y^H\;\mathrm{Cov}(X)\,x} \quad \text{and} \quad x^H\;\mathrm{Cov}(X)\,x \ge 0 .$$

The finite dimensional case obviously boils down to matrix calculation. Take X and Y $\mathbb{C}^d$ and $\mathbb{C}^\ell$ valued, respectively. Identifying linear operator from $\mathbb{C}^\ell$ to $\mathbb{C}^d$ as $d\times\ell$ matrices, we re-obtain the usual definition of covariance matrices

$$\mathrm{Cov}\,(X,Y) = [\mathrm{Cov}\,(X_i,Y_j)]_{\substack{1\le i\le d\\ 1\le j\le \ell}} = \mathbb{E}\left[XY^H\right] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y]^H .$$

Recall the useful identity in the finite-dimensional case, for any matrices A, B with d and ℓ columns, respectively, we have

$$\mathrm{Cov}\,(AX, VY) = A\,\mathrm{Cov}\,(X,Y)\;B^H .$$

The same identity obviously hold in the general case.

Proposition 6.2.1. *Let $\mathcal{H}_0$ and $\mathcal{G}_0$ be two Hilbert spaces. Suppose that X and Y are r.v. in $\mathcal{H} = L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} = L^2(\Omega, \mathcal{G}_0, \mathcal{F}, \mathbb{P})$, respectively. Let $\mathcal{H}_0'$ and $\mathcal{G}_0'$ be two Hilbert spaces and let $A \in \mathcal{L}_b(\mathcal{H}_0, \mathcal{H}_0')$ and $B \in \mathcal{L}_b(\mathcal{G}_0, \mathcal{G}_0')$. Then AX and BY are in $\mathcal{H}' = L^2(\Omega, \mathcal{H}_0', \mathcal{F}, \mathbb{P})$ and $\mathcal{G}' = L^2(\Omega, \mathcal{G}_0', \mathcal{F}, \mathbb{P})$, respectively, and*

$$\operatorname{Cov}(AX, BY) = A \operatorname{Cov}(X, Y)\, B^H ,$$

where B^H is the adjoint operator of B defined on $\mathcal{G}_0$ by

$$\left\langle B^H(x), y \right\rangle_{\mathcal{G}_0} = \langle x, B(y) \rangle_{\mathcal{G}_0} , \qquad y \in \mathcal{G}_0 .$$

It is well known that every quadratic form admits a *polarization identity* which allows one to recover its associated sesquilinear form. So if A is a Hermitian operator ($A^H = A$) on $\mathcal{H}_0$, then $(x, y) \mapsto y^H A\, x$ is its associated sesquilinear form and $Q : x \mapsto x^H A\, x$ its associated quadratic form, and we have, for all $x, y \in \mathcal{H}_0$,

$$y^H A\, x = \frac{1}{4} \left(Q(x + y) - Q(x - y) + \mathrm{i}\, Q(x + \mathrm{i}y) - \mathrm{i}\, Q(x - \mathrm{i}y) \right) . \tag{6.3}$$

In fact, this identity does not require A to be Hermitian. That it, for any linear operator A from $\mathcal{H}_0$ to itself, one can define a mapping $Q : x \mapsto x^H A\, x$ (which is called a quadratic form if and only if it is real valued, which is the case if and only if A is Hermitian) and the identity (6.3) holds.

6.3 Weakly stationary time series in Hilbert space

Having defined an expectation and a covariance for random variables, we easily extend the definition of weakly stationary time series valued in $\mathbb{C}$ (see Definition 2.3.1) to time series valued in any Hilbert space $\mathcal{H}_0$.

Definition 6.3.1 (Weakly stationary time series valued in a Hilbert space)**.** *Let $X = (X_t)_{t \in \mathbb{Z}}$ be an L^2 process valued in a Hilbert space $\mathcal{H}_0$. We say that X is weakly stationary if there exists $\mu \in \mathcal{H}_0$ and $\Gamma : \mathbb{Z} \to \mathcal{L}_b(\mathcal{H}_0)$ such that*

(i) For all $t \in \mathbb{Z}$, $\mathbb{E}[X_t] = \mu$.

(ii) For all $s, t \in \mathbb{Z}$, $\operatorname{Cov}(X_s, X_t) = \Gamma(s - t)$.

We call μ the mean *of X and Γ the* covariance operator function*.*

Recall that, in the univariate case, by the Cauchy-Schwartz inequality the covariance function modulus is at most the variance. The somewhat equivalent inequality holds in the multivariate case. Namely, as a consequence of (6.2), we have, for all $\tau \in \mathbb{Z}$,

$$\|\Gamma(\tau)\| \leq \operatorname{Var}(X) . \tag{6.4}$$

Obviously, we can extract various univariate weakly time series from one valued ion a Hilbert space, by applying a constant linear form.

Proposition 6.3.1. *Let $X = (X_t)_{t \in \mathbb{Z}}$ be a L^2 process valued in a Hilbert space $\mathcal{H}_0$ and let $A \in \mathcal{L}_b(\mathcal{H}_0, \mathcal{G}_0)$ for some Hilbert space $\mathcal{G}_0$. Suppose that X is weakly stationary with mean μ and covariance operator function Γ. Then $(A\, X_t)_{t \in \mathbb{Z}}$ is weakly stationary with mean $A\, \mu$ and covariance operator function $A\, \Gamma\, A^H$.*

In particular, for all $x \in \mathcal{H}_0$, $\left(\langle X_t, x\rangle_{\mathcal{H}_0}\right)_{t\in\mathbb{Z}}$ is weakly stationary with mean $\langle \mu, x\rangle_{\mathcal{H}_0}$ and covariance function $t \mapsto x^H \Gamma(t)\, x$. On the other hand using the polarization formula (6.3), one can deduce the operator $\Gamma(t)$ from $x \mapsto x^H \Gamma(t)\, x$. Hence we have the following.

Proposition 6.3.2. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a L^2 process valued in a Hilbert space $\mathcal{H}_0$. Then X is weakly stationary if and only if, for all $x \in \mathcal{H}_0$, $\langle X, x\rangle_{\mathcal{H}_0}$ is weakly stationary.*

We can also use this trick to rely on the univariate case for defining a spectral measure to represent the covariance structure of a weakly stationary process valued in a Hilbert space. We differ this to Section 6.6, where we restrict ourselves to the finite dimensional case in order to avoid too advanced operator theory on Hilbert spaces.

In the multivariate case, $\mathcal{H}_0 = \mathbb{C}^d$ for some $d \geq 1$, Proposition 6.3.1 just means that, for any matrix A with d columns, if $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ is weakly stationary, so is $\mathbf{Y} = (A\,\mathbf{X}_t)_{t\in\mathbb{Z}}$. This process is called an *instantaneous mixture* of $\mathbf{X}$, since at each time instant, the entries of $\mathbf{Y}_t$ are obtained as linear combinations of the entries of $\mathbf{X}_t$. The next proposition allows one to introduce *convolution mixtures*.

Proposition 6.3.3. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be an L^2 process valued in a Hilbert space $\mathcal{H}_0$. Let $\mathcal{G}_0$ be another Hilbert space and let $A = (A_t)_{t\in\mathbb{Z}} \in \mathcal{L}_b(\mathcal{H}_0, \mathcal{G}_0)^{\mathbb{Z}}$ satisfying the absolute summability condition*

$$\sum_{t\in\mathbb{Z}} \|A_t\| < \infty\ . \tag{6.5}$$

Suppose that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left[|X_t|^2_{\mathcal{H}_0}\right] < \infty\ . \tag{6.6}$$

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

$$Y_t = \sum_{s\in\mathbb{Z}} A_s X_{t-s}\ , \tag{6.7}$$

is absolutely convergent in $\mathcal{G}_0$ a.s. and in $\mathcal{G} = L^2(\Omega, \mathcal{G}_0, \mathcal{F}, \mathbb{P})$. Moreover, if X is weakly stationary with mean μ and covariance operator function Γ, then $Y = (Y_t)_{t\in\mathbb{Z}}$ is stationary with mean

$$\mu' = \sum_{s\in\mathbb{Z}} A_s\, \mu$$

and covariance operator function

$$\Gamma'(\tau) = \sum_{s,u\in\mathbb{Z}} A_s\, \Gamma(\tau + u - s)\, A_u^H\ , \qquad \tau \in \mathbb{Z}\ . \tag{6.8}$$

Proof. The proof is exactly the same one as in the univariate case, see Section 3.1. Following this proof, the infinite sum in (6.8) is defined by setting, for all $\tau \in \mathbb{Z}$ and $x, y \in \mathcal{G}_0$,

$$x^H\, \Gamma'(\tau)\, y = \sum_{s,u\in\mathbb{Z}} x^H\, A_s\, \Gamma(\tau + u - s)\, A_u^H\, y$$

which we know should converge in $\mathcal{G}$ (by continuity of its inner product). Note also that the infinite sum in (6.8) can be shown directly to converge absolutely in $\mathcal{L}_b(\mathcal{G}_0)$. Indeed, by (6.4), we have for all $\tau, s, u \in \mathbb{Z}$,

$$\left\|A_s\, \Gamma(\tau + u - s)\, A_u^H\right\| \leq \|A_s\|\ \|\Gamma(\tau + u - s)\|\ \left\|A_u^H\right\| \leq \|A_s\|\ \|A_u\|\ \mathrm{Var}\,(X)\ ,$$

which is summable over $u, s \in \mathbb{Z}$. □

Remark 6.3.1. *In Proposition 6.3.4, Condition (6.6) is always assumed. A natural question is to ask whether it automatically holds in the case where X is weakly stationary. In this case we know by assumption that $\mathbb{E}\left[|X_t|^2_{\mathcal{H}_0}\right] < \infty$ for all t, and, having in mind the finite dimensional case, we expect this quantity to depend only on $\mu = \mathbb{E}[X_t]$ and $\Gamma(0) = \mathrm{Cov}(X_t)$, which do not depend on t, which would immediately yields (6.6). It remains indeed true in the infinite dimensional case, at least if $\mathcal{H}_0$ is separable. To see why, take a Hilbert basis $(\phi_n)_{n\in\mathbb{N}}$ of $\mathcal{H}_0$. We can then write*

$$\mathbb{E}\left[|X_t|^2_{\mathcal{H}_0}\right] = \sum_{n\in\mathbb{N}} \mathbb{E}\left[\left|\langle X_t, \phi_n\rangle_{\mathcal{H}_0}\right|^2\right] = \sum_{n\in\mathbb{N}} \phi_n^H\, \Gamma(0)\, \phi_n + \sum_{n\in\mathbb{N}} \left|\langle \mu, \phi_n\rangle_{\mathcal{H}_0}\right|^2 ,$$

which are two infinite sums of non-negative summands which do not depend on t. Hence finiteness for all (even one) $t \in \mathbb{Z}$ implies finiteness of the sup over $t \in \mathbb{Z}$. Note also that, in the right-hand side of the previous display, the second term is simply $|\mu|^2_{\mathcal{H}_0}$. The first term is called the trace *of $\Gamma(0)$ (see Section 6.4).*

In the following, we use similar notation as in the univariate case, namely the process Y defined by (6.7) will be denoted as

$$Y = \mathrm{F}_A(X) .$$

Note also that if A is finitely supported we can write

$$\mathrm{F}_A(X) = \sum_{s\in\mathbb{Z}} A_s\, \mathrm{B}^s(X) ,$$

a notation which is often used when A is not finitely supported. When $\mathcal{H}_0$ and $\mathcal{G}_0$ have finite dimensions (so A is a sequence of matrices) the process Y is called a *convolution mixture* of X. As in the univariate case, we easily derive the following.

Proposition 6.3.4. *Let $\mathcal{G}_0, \mathcal{H}_0, \mathcal{I}_0$ be Hilbert spaces and let $A = (A_t)_{t\in\mathbb{Z}} \in \mathcal{L}_b(\mathcal{H}_0, \mathcal{G}_0)^{\mathbb{Z}}$ and $B = (B_t)_{t\in\mathbb{Z}} \in \mathcal{L}_b(\mathcal{G}_0, \mathcal{I}_0)^{\mathbb{Z}}$ be absolutely summable. Then, for any weakly stationary process $X = (X_t)_{t\in\mathbb{Z}}$ valued in $\mathcal{H}_0$, we have*

$$\mathrm{F}_B \circ \mathrm{F}_A(X) = \mathrm{F}_{B\star A}(X) ,$$

where $B \star A$ is the sequence in $\mathcal{L}_b(\mathcal{H}_0, \mathcal{I}_0)^{\mathbb{Z}}$ defined by

$$[B \star A]_k = \sum_{j\in\mathbb{Z}} B_j\, A_{k-j} = \sum_{j\in\mathbb{Z}} B_{k-j}\, A_j .$$

6.4 A bit of operator theory

Suppose that $\mathcal{H}_0$ and $\mathcal{G}_0$ are separable Hilbert spaces. Then they admit Hilbert bases $(\phi_n)_{n\in\mathbb{N}}$ and $(\psi_n)_{n\in\mathbb{N}}$. Let $X \in \mathcal{H} = L^2(\Omega, \mathcal{H}_0, \mathcal{F}, \mathbb{P})$, $Y \in \mathcal{G} = L^2(\Omega, \mathcal{G}_0, \mathcal{F}, \mathbb{P})$ and suppose they are centered. Denote their covariance operator by

$$\Gamma = \mathrm{Cov}\,(X, Y) .$$

Then we have, for all $n, m \in \mathbb{N}$,

$$\langle \Gamma\psi_n, \phi_m\rangle_{\mathcal{H}} = \mathrm{Cov}\left(\langle X, \phi_m\rangle_{\mathcal{H}_0}, \langle Y, \psi_n\rangle_{\mathcal{G}_0}\right)$$

and thus

$$\begin{aligned}|\Gamma\psi_n|^2_{\mathcal{H}_0} &= \sum_{m\in\mathbb{N}} \left|\langle \Gamma\psi_n, \phi_m\rangle_{\mathcal{H}_0}\right|^2 \\ &\leq \sum_{m\in\mathbb{N}} \mathrm{Var}\left(\langle X, \phi_m\rangle_{\mathcal{H}_0}\right) \mathrm{Var}\left(\langle Y, \psi_n\rangle_{\mathcal{G}_0}\right) \\ &= \mathrm{Var}\left(\langle X, \phi_m\rangle_{\mathcal{H}_0}\right) \mathbb{E}\left[|Y|^2_{\mathcal{G}_0}\right] \ .\end{aligned}$$

Hence we obtain that

$$\sum_{n\in\mathbb{N}} |\Gamma\psi_n|^2_{\mathcal{H}_0} \leq \mathbb{E}\left[|X|^2_{\mathcal{H}_0}\right] \mathbb{E}\left[|Y|^2_{\mathcal{G}_0}\right] \ .$$

By [Young, 1988, Defintion 8.5], we get that the covariance operator is *Hilbert-Schmidt* hence a *compact operator*. In the case $X = Y$, we already noticed that Γ is moreover non-negative definite Hermitian.

One can develop a singular value decomposition theory for these operators. A non-negative definite Hermitian Hilbert-Schmidt operator is called a *trace-class operator* if moreover the sum of all its eigen values is finite. Then its trace is defined as this sum. In the case of a, not Hermitian, Hilbert-Schmidt operator Γ from $\mathcal{H}_0$ to itself, it is called a *trace-class operator* if the non-negative definite Hermitian Hilbert-Schmidt operator $\left(\Gamma^H\Gamma\right)^{1/2}$ is a trace-class operator. In all these cases, the trace of Γ can be computed by the formula

$$\mathrm{Trace}\,\Gamma = \sum_{n\in\mathbb{N}} \langle \Gamma\psi_n, \psi_n\rangle_{\mathcal{H}_0} \ .$$

Note that in the case $\mathcal{H}_0 = \mathcal{G}_0$ and $\Gamma = \mathrm{Cov}\,(X, Y)$ as above, we have

$$\begin{aligned}\sum_{n\in\mathbb{N}} \langle \Gamma\psi_n, \psi_n\rangle_{\mathcal{H}_0} &= \sum_{n\in\mathbb{N}} \mathrm{Cov}\left(\langle X, \psi_n\rangle_{\mathcal{H}_0}, \langle Y, \psi_n\rangle_{\mathcal{H}_0}\right) \\ &= \sum_{n\in\mathbb{N}} \mathbb{E}\left[\langle X, \psi_n\rangle_{\mathcal{H}_0} \overline{\langle Y, \psi_n\rangle_{\mathcal{H}_0}}\right] \\ &= \mathbb{E}\left[\sum_{n\in\mathbb{N}} \langle X, \psi_n\rangle_{\mathcal{H}_0} \overline{\langle Y, \psi_n\rangle_{\mathcal{H}_0}}\right] \\ &= \mathbb{E}\left[\langle X, Y\rangle_{\mathcal{H}_0}\right] \ .\end{aligned}$$

The inversion of $\sum_{n\in\mathbb{N}}$ and $\mathbb{E}$ that we just applied is justified by

$$\begin{aligned}\mathbb{E}\left[\sum_{n\in\mathbb{N}} \left|\langle X, \psi_n\rangle_{\mathcal{H}_0} \overline{\langle Y, \psi_n\rangle_{\mathcal{H}_0}}\right|\right] &\leq \mathbb{E}\left[\left(\sum_{n\in\mathbb{N}} \left|\langle X, \psi_n\rangle_{\mathcal{H}_0}\right|^2 \sum_{n\in\mathbb{N}} \left|\langle Y, \psi_n\rangle_{\mathcal{H}_0}\right|^2\right)^{1/2}\right] \\ &= \mathbb{E}\left[|X|_{\mathcal{H}_0} |Y|_{\mathcal{H}_0}\right] \\ &\leq \left(\mathbb{E}\left[|X|^2_{\mathcal{H}_0}\right] \mathbb{E}\left[|Y|^2_{\mathcal{H}_0}\right]\right)^{1/2} < \infty \ .\end{aligned}$$

In particular in the case $X = Y$ we obtain the identity, well known when X is a centered L^2 $\mathbb{C}^d$-valued r.v.,

$$\mathrm{Trace}\ \mathrm{Cov}(X) = \mathbb{E}\left[|X|^2_{\mathcal{H}_0}\right] \ .$$

6.5 Finite dimensional case: multivariate time series

In the case where $\mathcal{H}_0 = \mathbb{C}^d$ with $d \in \mathbb{N}^*$, a process $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ valued in $\mathcal{H}_0$ is called a *multivariate time series*. Additional definitions are often used in this case, eased by the fact that continuous linear operators from $\mathcal{H}_0$ to itself are identified to $d \times d$ matrices. Suppose that $\mathbf{X}$ is weakly stationary with covariance matrix function Γ. Denote by D the diagonal matrix whose diagonal coincides with that of $\Gamma(0)$. In other words D contains the variances of the entries of the random vectors $\mathbf{X}_t$. Hence we can assume that the diagonal entries of D are non-zero or otherwise just discard the entries of $\mathbf{X}$ with zero variances to obtain such a D. Then the matrix

$$R(0) = D^{-1/2}\Gamma(0)D^{-1/2}$$

is called the *correlation matrix* of $\mathbf{X}$. It is a matrix of correlations, hence takes its values in $[-1, 1]$. More generally, the correlation matrix function is defined as

$$R(t) = D^{-1/2}\Gamma(t)D^{-1/2}\,, \qquad t \in \mathbb{Z}\,,$$

and also is a matrix of correlations for all $t \in \mathbb{Z}$.

6.6 Spectral representations of multivariate time series

In this section, we consider for simplicity the finite dimensional case $\mathcal{H}_0 = \mathbb{C}^d$ with $d \in \mathbb{N}^*$. The case $d \geq 2$ is often refer to as *multivariate time series*. We call an operator (or matrix) measure ν on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ a collection $(\nu_{k,\ell})_{1\leq k,\ell\leq d}$ of finite complex valued measures on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. For all x, y defined on $\mathbb{T}$ and taking values in $\mathcal{H}_0$, we use the notation

$$\int_{\mathbb{T}} x^H(\lambda)\nu(\mathrm{d}\lambda)y(\lambda) = \int_{\mathbb{T}} \sum_{k,l} \overline{x_k(\lambda)}\, \nu_{k,l}(\mathrm{d}\lambda)\, y_l(\lambda) \ .$$

Observe that, in particular, for given $x, y \in \mathcal{H}_0$, the measure $x^H \nu\, y$ is defined by

$$\left[x^H \nu\, y\right](A) = \int_{\mathbb{T}} \mathbb{1}_A(\lambda)\; x^H \nu(\mathrm{d}\lambda)\, y\,, \qquad A \in \mathcal{B}(\mathbb{T})\,.$$

Moreover the operator measure ν is completely determined by the collection of complex measures $\left(x^H \nu\, y\right)_{x,y\in\mathcal{H}_0}$, and thus, by the polarization formula (6.3), also by $\left(x^H \nu\, x\right)_{x\in\mathcal{H}_0}$.

Definition 6.6.1 (Non-negative definite operator measures)**.** *We say that the operator measure ν is* non-negative definite *if for all $x \in \mathrm{C}_b(\mathbb{T}, \mathcal{H}_0)$,*

$$\int_{\mathbb{T}} x^H(\lambda)\, \nu(\mathrm{d}\lambda)\, x(\lambda) \in \mathbb{R}_+\,.$$

Using the general construction of (complex valued) Gaussian processes (see Section 1.3) there exists a weakly stationary process $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ valued in $\mathcal{H}_0$ with covariance operator function Γ, if and only if, $t \mapsto \Gamma(t)$ is *non-negative definite*, which, in the multivariate case means that, for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{H}_0$, if Γ is given by (6.9), then

$$\sum_{s,t=1}^{n} x_s^H \Gamma(s-t) x_t \geq 0\,.$$

Theorem 6.6.1 (Spectral measure of a multivariate time series)**.** *Let $d \geq 1$ and denote $\mathcal{H}_0 = \mathbb{C}^d$. A function $\Gamma : \mathbb{Z} \to \mathcal{L}_b(\mathcal{H}_0)$ is* non-negative definite *if and only if there exists a finite Hermitian non-negative $d \times d$ matrix measure ν on $\mathbb{T}$ such that, for all $t \in \mathbb{Z}$,*

$$\Gamma(t) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}\lambda t}\, \nu(\mathrm{d}\lambda)\ . \tag{6.9}$$

Moreover ν is uniquely defined by the identity (6.9) for all $t \in \mathbb{Z}$ and, if $\mathbf{X}$ is weakly stationary with autocovariance operator Γ, the so defined measure ν is called the spectral measure matrix *of $\mathbf{X}$. If ν admits a density $f : \mathbb{T} \to \mathbb{C}^{d\times d}$ with respect to the Lebesgue measure, then f is called the* spectral density matrix *of $\mathbf{X}$.*

Proof. We rely on the proof of the Herglotz theorem, see Theorem 2.4.1. The *if* part is proved similarly.

The *only if* part requires to construct the matrix measure ν. Denote, for all $k \in \mathbb{Z}$,

$$\widehat{\Gamma}_k(\lambda) = \frac{1}{2\pi} \sum_{j=-k}^{k} \Gamma(j)\, \mathrm{e}^{-\mathrm{i}\lambda j}\ ,$$

and its Cesaro sum

$$F_n(\lambda) = \frac{1}{n} \sum_{k=0}^{n-1} \widehat{\Gamma}_k(\lambda) = \frac{1}{2\pi n} \sum_{\ell=1}^{n} \sum_{m=1}^{n} \Gamma(\ell - m)\, \mathrm{e}^{-\mathrm{i}\lambda(\ell-m)}\ . \tag{6.10}$$

Then the fact that $t \mapsto \Gamma(t)$ is non-negative definite yields, for all $x \in \mathcal{H}_0$ and all $\lambda \in \mathbb{T}$,

$$x^H\, F_n(\lambda)\, x = \frac{1}{2\pi n} \sum_{\ell=1}^{n} \sum_{m=1}^{n} (\mathrm{e}^{\mathrm{i}\lambda\ell}\, x)^H\, \Gamma(\ell - m)\, (\mathrm{e}^{-\mathrm{i}\lambda m}\, x) \geq 0\ .$$

Moreover we have, for all $t \in \mathbb{Z}$ and all $x, y \in \mathcal{H}_0$,

$$\lim_{n\to\infty} \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}\lambda t}\, x^H\, F_n(\lambda)\, y\, \mathrm{d}\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j=-k}^{k} x^H\, \Gamma(j)\, y\, \mathbb{1}_{\{j=t\}} = x^H\, \Gamma(t)\, y\ . \tag{6.11}$$

The idea is to define ν as a limit of the multivariate measure with density F_n on $\mathbb{T}$. For this we follow the lines of the proof of Theorem 2.4.1, and rely on the Prohorov Theorem for univariate probability measures. Take $x \in \mathcal{H}_0$. If $x^H\Gamma(0)x = 0$, then, by Cauchy-Schwartz inequality, $x^H\Gamma(t)x = 0$ for all $t \in \mathbb{Z}$ and $F_n = 0$ for all $n \in \mathbb{N}$. If $x^H\Gamma(0)x > 0$, then, using Prohorov's theorem, the probability measure with density $\lambda \mapsto x^H F_n(\lambda)x / x^H\Gamma(0)x$ on $\mathbb{T}$ converges weakly to the probability measure P_x *uniquely* defined by

$$\int_{\mathbb{T}} \mathrm{e}^{\lambda t} P_x(\mathrm{d}\lambda) = \frac{x^H\Gamma(t)x}{x^H\Gamma(0)x} \quad \text{for all } t \in \mathbb{Z}\ .$$

Hence we get that, for all $g \in \mathrm{C}_b(\mathbb{T})$ and $x \in \mathcal{H}_0$,

$$\lim_{n\to\infty} \int_{\mathbb{T}} g(\lambda)\, x^H\, F_n(\lambda)\, x\, \mathrm{d}\lambda = \int_{\mathbb{T}} g\, \mathrm{d}\nu_x\ , \tag{6.12}$$

where ν_x is the zero measure if $x^H\Gamma(0)x = 0$ and is the finite non-negative measure $x^H\Gamma(0)x\, P_x$ otherwise. In particular, using (6.11), we have, for all $t \in \mathbb{Z}$,

$$\int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}\lambda t}\, \mathrm{d}\nu_x = x^H\, \Gamma(t)\, x\ , \tag{6.13}$$

and this identity uniquely defines the non-negative measure ν_x for all $x \in \mathcal{H}_0$.

The operator measure ν is then defined using the polarization formula (6.3), that is, for all $x, y \in \mathcal{H}_0$,

$$x^H\, \nu(\mathrm{d}\lambda)\, y = \frac{1}{4}\left(\nu_{x+y}(\mathrm{d}\lambda) - \nu_{x-y}(\mathrm{d}\lambda) + \mathrm{i}\,\nu_{x+\mathrm{i}y}(\mathrm{d}\lambda) - \mathrm{i}\,\nu_{x-\mathrm{i}y}(\mathrm{d}\lambda)\right)\ ,$$

Applying the same polarization formula to $\Gamma(t)$, we get Relation (6.9). The fact that the obtained ν is non-negative definite also follows from the polarization formula, this time with (6.12), which yields, for all $x, y \in \mathrm{C}_b(\mathbb{T}, \mathcal{H}_0)$,

$$\lim_{n\to\infty} \int_{\mathbb{T}} x^H(\lambda)\ F_n(\lambda)\, y(\lambda)\ \mathrm{d}\lambda = \int_{\mathbb{T}} x^H(\lambda)\, \nu(\mathrm{d}\lambda)\, y(\lambda)\ .$$

In particular, for $x = y$, since $x^H(\lambda)F_n(\lambda)x(\lambda) \geq 0$ for all $\lambda \in \mathbb{T}$, we get that ν is non-negative definite. □

Observe that if a weakly stationary multivariate time series $\mathbf{X}$ admits a spectral density matrix f, then the diagonal of f contains the spectral density functions of the components of $\mathbf{X}$. In fact, it suffices that all components of $\mathbf{X}$ admit a spectral density for the process $\mathbf{X}$ to admit one as well. (See Exercise 6.3).

The off-diagonal components of f are called the *cross-spectral* density functions. The *coherence function* between two components k and ℓ is defined as the normalized square modulus of cross-spectral density

$$C_{k,\ell}(\lambda) := \frac{|f_{k,\ell}(\lambda)|^2}{f_{k,k}(\lambda) f_{\ell,\ell}(\lambda)}\ , \qquad \lambda \in \mathbb{T}\ ,$$

with the convention $0/0 = 0$. It is valued in $[0, 1]$, see Exercise 6.3.

Example 6.6.1 (Multivariate white noise)**.** *Let $d \geq 1$ and Σ be a $d \times d$ non-negative definite Hermitian matrix. We write $\mathbf{Z} \sim \mathrm{WN}(\mu, \Sigma)$ and say that $\mathbf{Z}$ is a* multivariate white noise *if it is a weakly stationary process with mean μ and covariance matrix function*

$$\Gamma(t) = \begin{cases} \Sigma & \text{if } t = 0, \\ 0 & \text{otherwise} \end{cases}$$

Thus a weakly stationary process is a white noise if and only if it admits a constant spectral matrix density, and its covariance Σ is identified by

$$f(\lambda) = \frac{1}{2\pi}\,\Sigma\ .$$

Note that in this definition, the process $\mathbf{Z}$ is uncorrelated in time but can be correlated in space *(from one entry to another).*

Non-constant spectral densities can easily be obtained by linear filtering a white noise. The covariance structure of the convolution filtering defined in Section 6.3, as in the univariate case, can easily be described in the spectral domain.

Proposition 6.6.2. *Let $d \geq 1$ and denote $\mathcal{H}_0 = \mathbb{C}^d$ and $\mathcal{G}_0 = \mathbb{C}^\ell$. Let $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ be a weakly stationary process valued in $\mathcal{H}_0$ with covariance operator function Γ and let $A = (A_t)_{t\in\mathbb{Z}}$ be a sequence of $\ell \times d$ matrices such that*

$$\sum_{t\in\mathbb{Z}} \|A_t\| < \infty \,. \tag{6.14}$$

Then $\mathbf{Y} = \mathrm{F}_A(\mathbf{X})$ is weakly stationary with spacetral matrix measure

$$\nu'(\mathrm{d}\lambda) = A^*(\lambda)\nu'(\mathrm{d}\lambda)A^{*H}(\lambda) \,,$$

where, for all $\lambda \in \mathbb{R}$, $A^*(\lambda)$ is the matrix defined by

$$A^*(\lambda) = \sum_{t\in\mathbb{Z}} A_t \, \mathrm{e}^{-\mathrm{i}\lambda t} \,.$$

We pursue with the extension of the univariate case and now consider the spectral representation of a centered d-variate time series $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ itself. Let ν be its spectral matrix measure and let us denote

$$\mathcal{H}^{\mathbf{X}} = \overline{\mathrm{Span}}\left(\langle \mathbf{X}_s, x\rangle_{\mathcal{H}_0} \;:\; s \in \mathbb{Z},\, x \in \mathcal{H}_0\right)$$

the L^2-closure of all finite linear combinations obtained from the entries of $\mathbf{X}$ at any time. Similarly, we denote

$$\mathcal{H}^{\mathbf{X}}_t = \overline{\mathrm{Span}}\left(\langle \mathbf{X}_s, x\rangle_{\mathcal{H}_0} \;:\; s \leq t,\, x \in \mathcal{H}_0\right)$$

We observe that for x^H, y^H defined on $\mathbb{T}$ and valued in $\mathcal{H}_0^H := \left\{u^H \;:\; u \in \mathcal{H}_0\right\}$,

$$(x^H, y^H) \mapsto \left\langle x^H, y^H \right\rangle_\nu := \int_{\mathbb{T}} x^H(\lambda)\, \nu(\mathrm{d}\lambda)\, (y^H(\lambda))^H$$

defines an inner product on the Hilbert space

$$\mathcal{H}_\nu = \left\{ x^H : \mathbb{T} \to \mathcal{H}_0^H \;:\; \int_{\mathbb{T}} x^H(\lambda)\, \nu(\mathrm{d}\lambda)\, x(\lambda) < \infty \right\} \,.$$

Then, as in the univariate case, observing that, for all $s, t \in \mathbb{Z}$ and $x, y \in \mathcal{H}_0$,

$$\mathrm{Cov}\left(\langle \mathbf{X}_s, x\rangle_{\mathcal{H}_0}, \langle \mathbf{X}_t, y\rangle_{\mathcal{H}_0}\right) = x^H\, \Gamma(s-t)\, y = \int \left(x^H \mathrm{e}^{\mathrm{i}\lambda s}\right)\, \nu(\mathrm{d}\lambda)\, \left(y^H \mathrm{e}^{\mathrm{i}\lambda t}\right)^H \,,$$

it can be shown that $\mathcal{H}^{\mathbf{X}}$ is isomorphic to $\mathcal{H}_\nu$ through a (unique) unitary operator that maps, for any $s \in \mathbb{Z}$ and $x \in \mathcal{H}_0$, the element $\langle \mathbf{X}_s, x\rangle_{\mathcal{H}_0}$ to $\lambda \mapsto x^H \mathrm{e}^{\mathrm{i}s\lambda}$. As in the univariate case we will use the *stochastic integral* notation for the unitary operator from $\mathcal{H}_\nu$ to $\mathcal{H}^{\mathbf{X}}$, that is we write that

$$x^H \mapsto \int_{\mathbb{T}} x^H(\lambda)\, \hat{\mathbf{X}}(\mathrm{d}\lambda) \,. \tag{6.15}$$

is a unitary operator from $\mathcal{H}_\nu$ to $\mathcal{H}^{\mathbf{X}}$ In particular, we have, for all $s \in \mathbb{Z}$ and $x \in \mathcal{H}_0$,

$$x^H \mathbf{X}_s = \langle \mathbf{X}_s, x \rangle_{\mathcal{H}_0} = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}s\lambda} x^H \hat{\mathbf{X}}(\mathrm{d}\lambda) \; ,$$

and, for all $x, y \in \mathcal{H}_\nu$,

$$\mathrm{Cov}\left(\int x(\lambda)^H \widehat{\mathbf{X}}(\mathrm{d}\lambda), \int y(\lambda)^H \widehat{\mathbf{X}}(\mathrm{d}\lambda) \right) = \int_{\mathbb{T}} x^H(\lambda)\, \nu(\mathrm{d}\lambda)\, y(\lambda) \; .$$

The notation (6.15) of course can be extended to matrix valued functions A on $\mathbb{T}$ by considering each row of A separately, yielding

$$\int_{\mathbb{T}} A(\lambda)\, \hat{\mathbf{X}}(\mathrm{d}\lambda) \; ,$$

which is well defined if each row of A belong to $\mathcal{H}_\nu$, that is, if

$$\int_{\mathbb{T}} \mathrm{Trace}\left(A(\lambda)\, \nu(\mathrm{d}\lambda)\, A^H(\lambda) \right) < \infty \; .$$

An important example is obtained in the case of a convolution filtering where $\mathbf{Y} = \mathrm{F}_A(\mathbf{X})$, with $A = (A_t)_{t\in\mathbb{Z}}$ a sequence of matrices satisfying (6.14). Then we have, for all $t \in \mathbb{Z}$,

$$\mathbf{Y}_t = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}t\lambda} A^*(\lambda)\, \hat{\mathbf{X}}(\mathrm{d}\lambda) \; .$$

Another way to put it a concise way is to write

$$\hat{\mathbf{Y}}(\mathrm{d}\lambda) = A^*(\lambda)\, \hat{\mathbf{X}}(\mathrm{d}\lambda) \; .$$

We will also use the notation

$$\mathbf{Y} = \widehat{\mathrm{F}}_{A^*}(\mathbf{X}) \; .$$

6.7 Granger causality

The following definition originates from ideas introduced in Granger [1969].

Definition 6.7.1 (Granger-causality in distribution)**.** *Consider two univariate time series $X = (X_t)_{t\in\mathbb{Z}}$ and $Y = (Y_t)_{t\in\mathbb{Z}}$ defined on the same probability space. Suppose that $\mathcal{G} = (\mathcal{G}_t)_{t\in\mathbb{Z}}$ is a filtration such that $(Y_t)_{t\in\mathbb{Z}}$ is adapted to $\left(\mathcal{G}_t \vee \mathcal{F}_t^X\right)_{t\in\mathbb{Z}}$. We say that X* $\mathcal{G}$-Granger-causes *Y (in distribution) if the distribution of Y_{t+1} given $\mathcal{G}_t \vee \mathcal{F}_t^X$ is different from that of Y_{t+1} given $\mathcal{G}_t$. We say that X* instantaneously $\mathcal{G}$-Granger-causes *Y (in distribution) if the distribution of Y_{t+1} given $\mathcal{G}_t \vee \mathcal{F}_{t+1}^X$ is different from that of Y_{t+1} given $\mathcal{G}_t$.*

In this definition $\mathcal{G}_t \vee \mathcal{F}_t^X$ represents all the information available at time t and $\mathcal{G}_t$ all the information available at time t without observing X. It is clear that the so defined causality depends on the choice of the filtration $\mathcal{G}$. This is why we mention it in this definition. If the context makes it clear what the choice of $\mathcal{G}$ is, it is not necessary to be mentioned and one just writes *X Granger-causes Y*.

In a Gaussian context the Granger-causality in distribution take the simpler following L^2 form, which will be the one adopted in the following, unless specified otherwise.

Definition 6.7.2 (Granger-causality in L^2)**.** *Consider two centered L^2 univariate time series $X = (X_t)_{t\in\mathbb{Z}}$ and $Y = (Y_t)_{t\in\mathbb{Z}}$ and a centered vector valued L^2 process $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}}$ defined on the same probability space. Suppose that $\mathcal{H}_t^Y \subset \mathcal{H}_t^{\mathbf{Z}} + \mathcal{H}_t^X$ for all $t \in \mathbb{Z}$. We say that X* $\mathbf{Z}$-Granger-causes *Y if*

$$\mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right)\right) < \mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}}\right)\right) . \tag{6.16}$$

We say that X instantaneously $\mathbf{Z}$-Granger-causes *Y if*

$$\mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_t^X\right)\right) < \mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right)\right) . \tag{6.17}$$

Note that we always have

$$\mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right)\right) \le \mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}}\right)\right) .$$

and, by Pythagoras' theorem, since

$$Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right) \perp \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right) - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}}\right) ,$$

the equality is equivalent to

$$\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right) = \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}}\right) .$$

Hence (6.16) is equivalent to saying that $\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right)$ and $\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}}\right)$ are not the same r.v.'s. And, similarly (6.17) is equivalent to saying that $\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_t^X\right)$ and $\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-1}^X\right)$ are not the same r.v.'s.

In Definition 6.7.2, saying that X Granger-causes Y means that the past of X up to time $t-1$ is useful to predict Y_t, and saying that X instantaneously Granger-causes Y means that the past of X up to time t is even more useful. If X does not instantaneously Granger-causes Y, it is useless to consider X_t to predict Y_t. But this might also be true about X_{t-1}, which may not improve the prediction based only on $\mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-2}^X$. This is the motivation for the following definition.

Definition 6.7.3 (Granger-causality lag)**.** *Consider the same setting as Definition 6.7.2. The* $\mathbf{Z}$-Granger-causality lag *of X to predict Y is defined as the smallest integer that belongs to*

$$\left\{k \ge 1 \,:\, \mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-k}^X\right)\right) < \mathrm{Var}\left(Y_t - \mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-k-1}^X\right)\right)\right\} .$$

If we include $k = 0$ in the above set, instantaneous causality would correspond to a Granger-causality lag equal to zero.

This definition means that if the Granger-causality lag is m, then X_{t-j}, $j = 0, \ldots, m-1$, do not improve the prediction of X_t by $\mathrm{proj}\left(Y_t \middle| \mathcal{H}_{t-1}^{\mathbf{Z}} + \mathcal{H}_{t-m}^X\right)$.

Example 6.7.1 (Noisy AR)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be an AR(1) process with AR coefficient $\phi \in (-1,1) \setminus \{0\}$ and innovation $\epsilon \sim \mathrm{WN}\left(0, \sigma^2\right)$, with $\sigma^2 > 0$, and $(Y_t)_{t\in\mathbb{Z}}$ be defined by*

$$Y_t = X_t + \eta_t, \qquad t \in \mathbb{Z} ,$$

where $\eta \sim \mathrm{WN}(0,1)$ is uncorrelated with ϵ. Taking $\mathbf{Z} = Y$, it is easy to show (see Exercise 6.9) that X both Granger-causes and instantaneously Granger-causes Y.

6.8 Vector ARMA processes

6.8.1 Reduced form representation

The natural multidimensional extension of ARMA process is given by the following definition.

Definition 6.8.1 (VARMA processes)**.** *Let $d \geq 2$, a $\mathbb{R}^d$-valued process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is called a vector autoregressive moving average or order p, q (VARMA(p, q)) process if it is weakly stationary and satisfies a VARMA equation, for all $t \in \mathbb{Z}$,*

$$\mathbf{X}_t = \sum_{k=1}^{p} \Phi_k \mathbf{X}_{t-k} + \mathbf{Z}_t + \sum_{k=1}^{q} \Theta_k \mathbf{Z}_{t-k} \ , \tag{6.18}$$

where $\Phi_1, \ldots, \Phi_p, \Theta_1, \ldots, \Theta_q$ are $d \times d$ matrices and $(\mathbf{Z}_t)_{t\in\mathbb{Z}}$ is a $\mathbb{R}^d$-valued white noise with covariance matrix Σ. In the case where $p = 0$, $\mathbf{X}$ is called a VMA(q) process and in the case where $q = 0$, it is called a VAR(p) process.

However, in the multivariate case, the study of the existence and uniqueness of stationary solutions to equation (6.18) is much more involved and, in most of the litterature, it often refer to *any process* satisfying the VARMA equation (6.18) in a *causal way*, that is, this equation is only verified for $t \in \mathbb{N}$ and in such a way that $\mathbf{X}_t$ belong to $\mathcal{H}_t^{\mathbf{Z}}$ for all $t \in \mathbb{N}$.

There is a simple sufficient condition that implies the existence of a causal stationary solution to the VARMA equation (6.18).

Theorem 6.8.1. *Suppose that the $d \times d$ matrices $\Phi_1, \ldots, \Phi_p$ are such that, for all $z \in \mathbb{C}$ such that $|z| \leq 1$, the matrix polyniomial*

$$\mathbf{\Phi}(z) := \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k$$

is invertible. Then there exists a unique stationary solution to (6.18) and this solution is causal, that is, for all $t \in \mathbb{Z}$,

$$\mathcal{H}_t^{\mathbf{X}} \subseteq \mathcal{H}_t^{\mathbf{Z}} \ .$$

If moreover the matrix polyniomial

$$\mathbf{\Theta}(z) = \mathbf{1} + \sum_{k=1}^{q} \Theta_k z^k$$

satisfies the same assumption as $\mathbf{\Phi}$, then the solution is also invertible *(hence* canonical*), that is, for all $t \in \mathbb{Z}$,*

$$\mathcal{H}_t^{\mathbf{Z}} = \mathcal{H}_t^{\mathbf{X}} \ .$$

Proof. Define the bloc companion matrix

$$\tilde{\Phi} = \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \ldots & \Phi_p \\ 1_d & 0 & 0 & \ldots & 0 \\ 0 & 1_d & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1_d & 0 \end{bmatrix} \ .$$

Then, defining the pd-dimenstional vectors

$$\tilde{\mathbf{X}}_t = \begin{bmatrix} \mathbf{X}_t \\ \vdots \\ \mathbf{X}_{t-p+1} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{Z}}_t = \begin{bmatrix} \mathbf{Z}_t + \sum_{k=1}^{q} \Theta_k \mathbf{Z}_{t-k} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Equation (6.18) is equivalent to the VAR(1) equation

$$\tilde{\mathbf{X}}_t = \tilde{\Phi}\tilde{\mathbf{X}}_{t-1} + \tilde{\mathbf{Z}}_t \ , \qquad t \in \mathbb{Z} \ .$$

Moreover, we have, for all $z \in \mathbb{C}$ and $\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_p \end{bmatrix} \in (\mathbb{R}^d)^p$,

$$\tilde{\Phi}\tilde{\mathbf{x}} = z\tilde{\mathbf{x}} \Longleftrightarrow \left(\sum_{k=1}^{p} \Phi_k z^{p-k} \right) \mathbf{x}_p = z^p \mathbf{x}_p \ .$$

Hence the condition on $\mathbf{\Phi}(z)$ guaranties that $\tilde{\Phi}$ has all its eigenvalues with moduli strictly less than 1. Hence there exists $c > 0$ and $\rho < 1$ such that, for all $k \in \mathbb{N}$,

$$\left| \tilde{\Phi}^k \right| \leq C \ \rho^k \ .$$

where $|\cdot|$ denotes the Euclidean operator norm on $\mathbb{R}^{p \times p}$. The conclusion of the proof is then similar to the VAR(1) case, as treated in the first questions of Exercise 6.5. □

The representation (6.18) is called a *reduced form* representation in the econometric literature as it provides a convenient and concise model for the whole vector $\mathbf{X}_t$ at once. Other representations are possible, and are investigated in Sections 6.8.2 and 6.8.3.

6.8.2 Impulse response (MA(∞) representation) and spectral matrix density

A completely different proof of Theorem 6.8.1 can be obtained by directly trying to inverse the operator as we did in the univariate case. Namely we look for a sequence $(A_s)_{s \in \mathbb{N}}$ of matrices such that

$$\sum_{s \in \mathbb{N}} \|A_s\| < \infty \quad \text{and} \quad \mathbf{\Phi}(\mathrm{B}) \circ \mathrm{F}_A = \mathrm{F}_A \circ \mathbf{\Phi}(\mathrm{B}) = \mathbf{1} \ ,$$

where the latter holds on the set of weakly stationary processes. This approach can be worked out by using the following result.

Lemma 6.8.2. *Let $\Phi_1, \dots, \Phi_p$ denote $d \times d$ complex matrices. Suppose that*

$$\rho := \sup \left\{ a > 0 \ : \ \forall z \in \mathbb{C}, |z| \leq a \Longrightarrow \mathbf{\Phi}(z) := \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k \ \textit{is invertible} \right\} > 0 \ .$$

Then there exists a sequence $A = (A_s)_{s \in \mathbb{N}^*}$ of $d \times d$ matrices such that, for all $a \in [0, \rho)$,

$$\sum_{s \in \mathbb{N}^*} \|A_s\| a^s < \infty \ ,$$

and, for all $z \in \mathbb{C}$ such that $|z| < \rho$, we have

$$\mathbf{\Phi}(z)^{-1} = \mathbf{1} + \sum_{s \geq 1} A_s \, z^s \; .$$

Moreover if $\Phi_1, \ldots, \Phi_p$ are real matrices, so are the matrices $A_1, A_2, \ldots$.

Proof. The inverse of an invertible matrix A can be computed using the determinant and the cofactor matrix, hence its entries are obtained by dividing a multivariate polynomial of the entries of A by an other non-vanishing multivariate polynomial of the entries A. Thus, since the composition and linear combinations of holomorphic functions is holomorphic, it follows from the assumption that all the entries of

$$z \mapsto \mathbf{\Phi}(z)^{-1}$$

are holomorphic on the complex open ball centered at the origin with radius ρ. It implies that they admit an entire series representation on this ball, which concludes the proof. □

The practical computation of the sequence A can be done by solving algebraically the equation

$$\left(\mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k \right) \left(\mathbf{1} + \sum_{s \geq 1} A_s \, z^s \right) = \mathbf{1} \; .$$

leading to $A_1 = \Phi_1$ and the recursive equations, for $j = 2, \ldots,$

$$A_j = \Phi_j + \sum_{k \geq 1, s \geq 1, s+k=j} \Phi_k A_j \; .$$

We deduce the following result.

Corollary 6.8.3. *Suppose that the $d \times d$ real matrices $\Phi_1, \ldots, \Phi_p$ are such that, for all $z \in \mathbb{C}$ such that $|z| \leq 1$, the matrix polyniomial*

$$\mathbf{\Phi}(z) = \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k$$

is invertible. Then there exists a causal sequence $A = (A_s)_{s \in \mathbb{Z}}$ of real $d \times d$ matrices with $A_0 = \mathbf{1}$ and $A_1 = \Phi_1$ such that, for any $\mathbb{R}^d$-valued weakly stationary process $(\mathbf{X}_t)_{t \in \mathbb{Z}}$,

$$\mathrm{F}_A \circ \mathbf{\Phi}(\mathrm{B})(X) = \mathbf{\Phi}(\mathrm{B}) \circ \mathrm{F}_A(\mathbf{X}) = \mathbf{X} \; .$$

Proof. We apply Lemma 6.8.2 in the case where $\rho > 1$. This gives that, defining the sequence B so that $\mathbf{\Phi}(\mathrm{B}) = \mathrm{F}_B$, we have $A \star B = B \star A$ equal to the impulse sequence. The result then follows from Proposition 6.3.4. □

Using this corollary, the proof of Theorem 6.8.1 is more straightforward and directly provide a representation of the VARMA process in the form

$$\mathbf{X}_t = \mathrm{F}_A \circ \Theta(\mathrm{B})(\mathbf{Z}) = \mathbf{Z}_t + \sum_{s \geq 1} \left(\sum_{k,l \geq 1, k+l=s} A_k \Theta_l + A_s + \Theta_s \mathbb{1}\{s \leq q\} \right) \mathbf{Z}_{t-s} \; .$$

Since $\mathbf{Z} \sim \mathrm{WN}(0, \Sigma)$, this can be seen as an $\mathrm{MA}(\infty)$ representation. An interesting consequence is that it provides the spectral density matrix of $\mathbf{X}$. Namely, applying Proposition 6.6.2, we get that $\mathbf{X}$ has spectral matrix density

$$f(\lambda) = \Phi(\mathrm{e}^{-\mathrm{i}\lambda})^{-1}\, \Theta(\mathrm{e}^{-\mathrm{i}\lambda})\, \frac{\Sigma}{2\pi}\, \Theta(\mathrm{e}^{-\mathrm{i}\lambda})^{H}\, \Phi(\mathrm{e}^{-\mathrm{i}\lambda})^{-1\,H} \,.$$

However in the econometric literature, an extra step is usually added, which consists in making the noise with uncorrelated entries. Namely, using a Choleskii factorization

$$L\,\Sigma\,L^{H} = \Delta \,. \tag{6.19}$$

of Σ, where L is a lower triangular matrix with unit entries on its diagonal, we obtain

$$\mathbf{X}_t = \mathrm{F}_A \circ \Theta(\mathrm{B})(\mathbf{Z}) = L^{-1}\tilde{\mathbf{Z}}_t + \sum_{s\geq 1} \left(\sum_{k,l\geq 1, k+l=s} A_k \Theta_l + A_s + \Theta_s \mathbb{1}\{s \leq q\} \right) L^{-1}\, \tilde{\mathbf{Z}}_{t-s} \,.$$

where $\tilde{\mathbf{Z}}_t = L\mathbf{Z}_t$ for all $t \in \mathbb{Z}$. Note that we computed explicitly the sequence B such that

$$\mathbf{X} = \mathrm{F}_B(\tilde{\mathbf{Z}}) \,.$$

This (causal) sequence B is called the *impulse response* of the VARMA process $\mathbf{X}$.

6.8.3 Structural form representation

We now introduce a new representation derived from the reduced form (6.18) which allows one to investigate instantaneous Granger causality. by splitting the entries of the vector $\mathbf{X}_t$ in order to make apparent the dependence structure within it. They read as, for all $t \in \mathbb{Z}$ and $\ell = 1, \ldots, d$,

$$\mathbf{X}_t(\ell) = \sum_{j=1}^{\ell-1} \psi_{\ell,j} \mathbf{X}_t(j) + \sum_{k=1}^{p} \left[\tilde{\Phi}_k \mathbf{X}_{t-k}\right](\ell) + \tilde{\mathbf{Z}}_t(\ell) + \sum_{k=1}^{q} \left[\tilde{\Theta}_k \tilde{\mathbf{Z}}_{t-k}\right](\ell) \,, \tag{6.20}$$

where $\tilde{\mathbf{Z}}$ is a white noise with *non-negative definite diagonal* covariance matrix Δ, and they are said to be in *structural form*. Equation (6.20) can be compactly written as

$$L\,\mathbf{X}_t = \sum_{k=1}^{p} \tilde{\Phi}_k \mathbf{X}_{t-k} + \tilde{\mathbf{Z}}_t + \sum_{k=1}^{q} \tilde{\Theta}_k \tilde{\mathbf{Z}}_{t-k} \,, \qquad t \in \mathbb{Z} \,, \tag{6.21}$$

where L is a lower triangular matrix with ones on its diagonal and $[-\psi_{\ell,j}]_{1\leq j<\ell\leq d}$ on its lower triangle. Hence we see that this form is easily obtained by setting

$$\tilde{\Phi}_k = L\,\Phi_k \,, \quad \tilde{\mathbf{Z}}_t = L\mathbf{Z}_t, \qquad \text{and} \quad \tilde{\Theta}_k = L\Theta_k L^{-1} \,, \tag{6.22}$$

and (6.19) allows us to recover the reduced form (6.18). Thus one can modify the representation back and forth from the reduced form (6.18) to the structural form (6.20) by choosing L lower triangular with ones on its diagonal so that the Choleski decomposition (6.19) holds.

This decomposition actually corresponds to the Gram Schimdt algorithm applied to the Hermitian form associated to the covariance matrix Σ :

$$\boldsymbol{\epsilon}(\ell) := \mathbf{Z}_0(\ell) - \mathrm{proj}\left(\mathbf{Z}_0(\ell) \middle| \mathrm{Span}\left(\mathbf{Z}_0(j),\, 1 \le j < \ell\right)\right) = \mathbf{Z}_0(\ell) - \sum_{j=1}^{\ell-1} \psi_{\ell,j} \mathbf{Z}_0(j)\ , \qquad 1 \le \ell \le d\ .$$

This indeed gives an orthogonal sequence $\boldsymbol{\epsilon}(\ell)$ for $\ell = 1, \dots, d$ and

$$\boldsymbol{\epsilon} = L\mathbf{Z}_0 \Longrightarrow \Delta = \mathrm{Cov}(\boldsymbol{\epsilon}) = L\Sigma L^H\ .$$

The matrices Δ and Σ are often assumed to be definite positive (which is equivalent to assuming that Σ is non-singular).

Remark 6.8.1. *It is important to note that the structural representation (6.20) relies on a particular ordering of the entries of the random vectors. Using this representation, each entries $\mathbf{X}_t(\ell)$ is expressed using the entries $\mathbf{X}_t(j)$ for $j < \ell$ and elements of $\mathcal{H}^{\mathbf{X}}_{t-1} + \mathcal{H}^{\mathbf{Z}}_{t-1}$ up to the additive error $\mathbf{Z}_t(\ell)$.*

We have seen in Theorem 6.8.1 sufficient conditions for defining a canonical VARMA process. As in the univariate case, canonical VARMA process provides a direct computation of predictors. We then have the following general result.

Theorem 6.8.4. *Let $d \ge 1$ and $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \Sigma)$ be a d-dimensional white noise. Suppose that $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ satisfies the VARMA equation (6.18) and that, for all $t \in \mathbb{Z}$,*

$$\mathcal{H}^{\mathbf{X}}_t = \mathcal{H}^{\mathbf{Z}}_t\ .$$

Then, for any $j \neq \ell \in \{1, \dots, d\}$, $\mathbf{X}(j)$ instantaneously $\mathbf{X}$-Granger causes $\mathbf{X}(\ell)$ if and only if $\mathbf{Z}(j)$ is correlated with $\mathbf{Z}(\ell)$.

Proof. We use the structural equation (6.20), which shows that, by a recursion on $\ell = 1, \dots, d$,

$$\mathcal{H}^{\mathbf{X}(\ell)}_t \subseteq \left(\mathcal{H}^{\mathbf{X}}_{t-1} + \mathrm{Span}\left(\mathbf{X}_t(j),\, 1 \le j < \ell\right)\right) \overset{\perp}{\oplus} \mathrm{Span}\left(\tilde{\mathbf{Z}}_t(\ell)\right)\ .$$

It also follows from this equation and the assumption that

$$\mathrm{proj}\left(\mathbf{X}_t(\ell) \middle| \mathcal{H}^{\mathbf{X}}_{t-1} + \mathrm{Span}\left(\mathbf{X}_t(j),\, 1 \le j < \ell\right)\right) = \sum_{j=1}^{\ell-1} \psi_{\ell,j}\mathbf{X}_t(j) + \sum_{k=1}^{p}\left[\tilde{\Phi}_k \mathbf{X}_{t-k}\right](\ell) + \sum_{k=1}^{q}\left[\tilde{\Theta}_k \tilde{\mathbf{Z}}_{t-k}\right](\ell)$$

and

$$\mathrm{proj}\left(\mathbf{X}_t(\ell) \middle| \mathcal{H}^{\mathbf{X}}_{t-1}\right) = \sum_{j=1}^{\ell-1} \psi_{\ell,j}\, \mathrm{proj}\left(\mathbf{X}_t(j) \middle| \mathcal{H}^{\mathbf{X}}_{t-1}\right) + \sum_{k=1}^{p}\left[\tilde{\Phi}_k \mathbf{X}_{t-k}\right](\ell) + \sum_{k=1}^{q}\left[\tilde{\Theta}_k \tilde{\mathbf{Z}}_{t-k}\right](\ell)\ .$$

In particular, we get that

$$\begin{aligned}\mathrm{proj}\left(\mathbf{X}_t(2) \middle| \mathcal{H}^{\mathbf{X}(\ell)}_{t-1} + \mathrm{Span}\left(\mathbf{X}_t(1)\right)\right) &= \mathrm{proj}\left(\mathbf{X}_t(2) \middle| \mathcal{H}^{\mathbf{X}(\ell)}_{t-1}\right) + \psi_{2,1}\left(\mathbf{X}_t(1) - \mathrm{proj}\left(\mathbf{X}_t(1) \middle| \mathcal{H}^{\mathbf{X}}_{t-1}\right)\right) \\ &= \mathrm{proj}\left(\mathbf{X}_t(2) \middle| \mathcal{H}^{\mathbf{X}(\ell)}_{t-1}\right) + \psi_{2,1}\tilde{\mathbf{Z}}_t(1)\ .\end{aligned}$$

Hence $\mathbf{X}(1)$ instantaneously $\mathbf{X}$-Granger causes $\mathbf{X}(2)$ if and only if $\psi_{2,1}\tilde{\mathbf{Z}}_t(1) \neq 0$. Since $\psi_{2,1}\tilde{\mathbf{Z}}_0(1) = \mathrm{proj}\left(\mathbf{Z}_0(2) \middle| \mathrm{Span}\left(\mathbf{Z}_0(1)\right)\right)$, we obtain the condition $\Sigma_{1,2} \neq 0$. □

VARMA models are not so much used for modeling time series (in contrast to VAR processes) or, when they are, practitioners use them with low orders. The basic reason is that they raise identifiability problems: a VARMA process can be the solution to to two different stable VARMA equations with the same noise! (which was not the case in the univariate case) We provide two examples hereafter. However, discarding this identifiability problem, (quasi-)maximum likelihood estimation is possible using the general framework of dynamic linear models as we will see in Chapter 7.

Example 6.8.1. *Consider the VAR(1) process* $\mathbf{X}$ *with AR matrix*

$$\Phi = \begin{bmatrix} 0 & \phi \\ 0 & 0 \end{bmatrix}$$

and white noise $\mathbf{Z}$*. Then it is a VMA(1) process with MA matrix*

$$\Theta = \begin{bmatrix} 0 & \phi \\ 0 & 0 \end{bmatrix}$$

and the same noise. See Exercise 6.7.

Example 6.8.2. *Consider the VARMA(1,1) process* $\mathbf{X}$ *with AR matrix*

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 + a \\ 0 & b \end{bmatrix}$$

and MA matrix

$$\Theta = \begin{bmatrix} \theta & -a \\ 0 & -b \end{bmatrix}$$

and white noise $\mathbf{Z}$*. Then* $\mathbf{X}$ *satisfies all the VARMA(1,1) equations obtained with these* Φ *and* Θ *and the same white noise, but with possibly different values of* a *and* b*. See Exercise 6.8.*

6.9 Exercises

Exercise 6.1. Let U and V be two centered and independent $\mathbb{C}$-valued random variables. We assume that U and V are L^2 and denote

$$\mathrm{Var}\,(U) = \sigma_1^2 \quad \text{and} \quad \mathrm{Var}\,(V) = \sigma_2^2\,,$$

which are both assumed to be strictly positive. We also define $W = U + V$ and denote

$$\gamma_1 = \mathrm{Cov}\left(U, \overline{U}\right) \quad \text{and} \quad \gamma_2 = \mathrm{Cov}\left(V, \overline{V}\right)\,,$$

where $\overline{U}$ and $\overline{V}$ are the complex conjugates of U and V, respectively. Recall that, since U and V are centered, we have $\sigma_1^2 = \mathbb{E}\left[|U|^2\right]$ and $\gamma_1 = \mathbb{E}\left[U^2\right]$, and similar formulas hold for V.

1. Find $\rho \in \mathbb{C}$ such that $\mathrm{Cov}\,(U - \rho W, W) = 0$.
2. Deduce an expression of $\hat{U}_1 = \mathrm{proj}\,(U|\,\mathrm{Span}\,(W))$.
3. Let $\hat{U}_2 = \mathrm{proj}\left(U|\,\mathrm{Span}\left(W, \overline{W}\right)\right)$. Give an expression of the matrix $A \in \mathbb{C}^{2\times 2}$ such that $\hat{U}_2 = \rho_1 W + \rho_2 \overline{W}$ with
$$A \begin{bmatrix}\rho_1 \\ \rho_2\end{bmatrix} = \begin{bmatrix}\sigma_1^2 \\ \gamma_1\end{bmatrix}\,.$$
4. Show that if $\gamma_1 = \sigma_1^2$ then there exists $\theta \in \mathbb{R}$, $\overline{U} = \mathrm{e}^{\mathrm{i}\theta} U$ a.s. What is θ if U is real a.s.? or purely imaginary a.s. ?

We say that a $\mathbb{C}$-valued variable T is circularly-symmetric if for all $\theta \in \mathbb{R}$, $\mathrm{e}^{\mathrm{i}\theta} T$ has the same distribution as T.

5. Let R be an $\mathbb{R}_+$-valued random variable and Φ a random variable with uniform distribution on $[0, 2\pi)$, independent of R. Show that $T = R\mathrm{e}^{\mathrm{i}\Phi}$ is circularly-symmetric.
6. Show that if T is an L^1 circularly-symmetric $\mathbb{C}$-valued variable, then it must be centered.
7. Show that if T is an L^2 circularly-symmetric $\mathbb{C}$-valued variable, then T^2 is an L^1 circularly-symmetric $\mathbb{C}$-valued variable. Deduce the value of $\mathrm{Cov}\left(T, \overline{T}\right)$ in this case.
8. Suppose that U and V are circularly-symmetric, L^2 and independent. Compare $\hat{U}_1$ and $\hat{U}_2$ defined above in this case.

Exercise 6.2 (Arbitrary covariance matrix). Let $p, q \geq 1$ and A be an arbitrary $p \times q$ complex matrix. Define, for any $t \in \mathbb{R}_+$,

$$M(t) = \begin{bmatrix}\mathbf{1}_p & t\,A \\ t\,A^H & \mathbf{1}_q\end{bmatrix}\,.$$

1. Show that $x^H\, M(t)\, x$ converges uniformly as $t \to 0$ on the set $\{x \in \mathbb{C}^{p+q} \,:\, |x| = 1\}$.
2. Show that there exists $t > 0$ small enough such that $M(t)$ is Hermitian non-negative semi-definite.
3. Deduce that there exists two random vectors X and Y such that $\mathrm{Cov}\,(X, Y) = A$. [**Hint**: use Section 1.3.]

Exercise 6.3. Let $\mathbf{X} = \left(\begin{bmatrix} X_t(1) & X_t(2) \end{bmatrix}^T \right)_{t\in\mathbb{Z}}$ be a weakly stationary bivariate time series with matrix spectral measure $\nu = [\nu_{j,k}]_{j,k=1,2}$. The goal of this exercise is to show that, if $X(1)$ and $X(2)$ admit spectral densities, then $\mathbf{X}$ admits a matrix spectral density $f = [f_{j,k}]_{j,k=1,2}$, and, for (Lebesgue-) a.e. $\lambda \in \mathbb{T}$, we have

$$|f_{1,2}(\lambda)|^2 \leq f_{1,1}(\lambda) f_{2,2}(\lambda) \; .$$

We assume without loss of generality that $\mathbf{X}$ is centered.

1. Show that, for all $A \in \mathcal{B}(\mathbb{T})$, we have

$$|\nu_{1,2}(A)|^2 \leq \nu_{1,1}(A) \; \nu_{2,2}(A) \; . \tag{6.23}$$

   [**Hint**: use the $\mathbb{C}^2$-valued random variable $\int \mathbb{1}_A \, \mathrm{d}\hat{\mathbf{X}}$.]

We now assume that $X(1)$ and $X(2)$ admit spectral densities.

2. Show that $\mathbf{X}$ admits a matrix spectral density $f = [f_{j,k}]_{j,k=1,2}$. [**Hint**: Recall the Hahn–Jordan decomposition of complex valued measure ξ as

$$\xi = \xi_+^r - \xi_-^r + \mathrm{i}\left(\xi_+^i - \xi_-^i\right) \; ,$$

   where ξ_+^r, ξ_-^r, ξ_+^i and ξ_-^i are nonnegative measure such that (ξ_+^r, ξ_-^r) and (ξ_+^i, ξ_-^i) are two pairs of singular measures. Now by the Radon-Nykodim thorem, we get that a finite complex valued measure ξ admits a density if and only if $\xi(A) = 0$ for all Borel set A with zero Lebesgue measure.]

3. Denoting

$$C_i = \{\lambda \in \mathbb{T} \, : \, |f_{1,2}(\lambda)| > 0 \text{ and } f_{i,i}(\lambda) = 0\} \; ,$$

   show that $\mathrm{Leb}(C_i) = 0$ for $i = 1, 2$.

For all $c \geq 0$, we define

$$A_c = \left\{\lambda \in \mathbb{T} \, : \, |f_{1,2}(\lambda)|^2 \geq c \; f_{1,1}(\lambda) f_{2,2}(\lambda) > 0\right\} \; ,$$

and, for all $0 \leq x < y$ and $i = 1, 2$,

$$B_{x,y}^{(i)} = \{\lambda \in \mathbb{T} \, : \, x \leq f_{i,i}(\lambda) \leq y\} \; .$$

4. Show that if $x_i \leq y_i$ for $i = 1, 2$ and $c x_1 x_2 > y_1 y_2$ then

$$\mathrm{Leb}\left(A_c \cap B_{x_1,y_1}^{(1)} \cap B_{x_2,y_2}^{(2)}\right) = 0 \; ,$$

   where Leb denotes the Lebesgue measure.

5. Deduce that $\mathrm{Leb}(A_c) = 0$ for all $c > 1$.

6. Conclude that $|f_{1,2}(\lambda)|^2 \leq f_{1,1}(\lambda) f_{2,2}(\lambda)$ for Leb-a.e. λ.

Exercise 6.4 (Vector MA(1)). Let $\epsilon \sim \mathrm{IID}(0, 1)$. Define $\mathbf{Z}_t = \begin{bmatrix} \epsilon_t & \epsilon_t \end{bmatrix}^T$. Define

$$\mathbf{X}_t = \mathbf{Z}_t + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{Z}_{t-1} \; .$$

1. Show that $\mathbf{X}$ is weakly stationary and compute its covariance operator function.
2. Determine its spectral density function.
3. Determine its coherence function $C_{1,2}$.
4. Show that $\mathbf{X}$ satisfies an MA(1) equation of the form
$$\mathbf{X}_t = \mathbf{x}\epsilon_t + \mathbf{y}\epsilon_{t-1} \ ,$$
where $\mathbf{x}$ and $\mathbf{y}$ are vectors of $\mathbb{R}^2$.
5. Use this equation to solve Questions 2 and 3.
6. Compute the best linear predictor of $\mathbf{X}_t$ given $\mathbf{X}_{t-1}$.

Exercise 6.5 (Vector AR(1)). Let $d \geq 1$. Consider the following vector AR(1) equation
$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{Z}_t \ , \tag{6.24}$$
where Φ is a $d \times d$ matrix, $(\mathbf{Z}_t)_{t\in\mathbb{Z}}$ is a centered process taking its values in $\mathbb{R}^d$ such that $\mathbb{E}[\mathbf{Z}_s \mathbf{Z}_t^T]$ vanishes if $s \neq t$ and equals $\sigma^2 I$ if $s = t$. The operator norm of Φ is denoted by
$$\|\Phi\| = \sup_{x\in\mathbb{R}^d, |x|\leq 1} |\Phi x| \ .$$
and the spectral radius of Φ by $\rho(\Phi)$. We assume that $\rho(\Phi) < 1$.

1. Show that there exists a unique weakly stationary vector process that satisfies (6.24).[**Indication**: use that $\|\Phi^k\| \leq C\rho^k$ for all $k \geq 1$, where $C > 0$ and $\rho(\Phi) < \rho < 1$ are some constants.]
2. Compute the autocovariance matrix function of $(\mathbf{X}_t)_{t\in\mathbb{Z}}$.
3. Compute the spectral density function of $(\mathbf{X}_t)_{t\in\mathbb{Z}}$.
4. Suppose that Φ is real symmetric with null space reduced to $\{0\}$. Write $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ as a linear transform of d uncorrelated scalar AR(1) processes $(Y_t(1))$,..., $(Y_t(d))$.
5. Let (V_t) denote the scalar process corresponding to the first coordinate of $(\mathbf{X}_t)_{t\in\mathbb{Z}}$. Which ARMA representation does it satisfy ?
6. Compute the spectral density of (V_t).

Exercise 6.6 (Cointegration). Consider a bivariate process $(\mathbf{X}_t)_{t\in\mathbb{N}}$ solution of the equation
$$\begin{bmatrix} \mathbf{X}_t(1) \\ \mathbf{X}_t(2) \end{bmatrix} = \underbrace{\begin{bmatrix} 1/2 & -1 \\ -1/4 & 1/2 \end{bmatrix}}_{\Phi} \begin{bmatrix} \mathbf{X}_{t-1}(1) \\ \mathbf{X}_{t-1}(2) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_t(1) \\ \boldsymbol{\epsilon}_t(2) \end{bmatrix} , \tag{6.25}$$
where $\boldsymbol{\epsilon}(1)$ and $\boldsymbol{\epsilon}(2)$ are two uncorrelated scalar white noise sequences with positive variances.

1. Show that Φ may be diagonalized as $P\Lambda P^{-1}$ where
$$P = \begin{pmatrix} 1 & 2 \\ -1/2 & 1 \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$
and compute P^{-1}.

2. Define $\mathbf{Y}_t = P^{-1}\mathbf{X}_t$. What kind of processes are $\mathbf{Y}(1)$ and $\mathbf{Y}(2)$? Is there a weakly stationary solution $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ to Equation (6.25)?
3. Show that there exists a unique linear combination of the form $\begin{bmatrix}1 & \alpha\end{bmatrix}\mathbf{X}_t$ that can be weakly stationary.

Exercise 6.7. Show that the claim in Example 6.8.1 is true.

Exercise 6.8. Consider $\mathbf{X}$ defined as in Example 6.8.2.

1. Give conditions on ϕ_1 and b to ensure that $\mathbf{X}$ is well defined.

We assume that the found conditions are satisfied in the following.

2. Show that, for all $t \in \mathbb{Z}$, we have
$$\mathbf{X}_t(2) - \mathbf{Z}_t(2) = b(\mathbf{X}_{t-1}(2) - \mathbf{Z}_{t-1}(2)) \ .$$
3. Show that it implies $\mathbf{X}(2) = \mathbf{Z}(2)$.
4. Deduce that the claim in Example 6.8.2 is true.

Exercise 6.9. Consider X and Y as in Example 6.7.1.

1. Show that Y is an ARMA(1,1) process and find its canonical representation and the variance v of its innovation, We denote by ζ the innovation process in the following.
2. Compute $\mathrm{proj}\left(Y_t | \mathcal{H}^Y_{t-1} + \mathcal{H}^X_t\right)$ and $\mathrm{proj}\left(Y_t | \mathcal{H}^Y_{t-1} + \mathcal{H}^X_{t-1}\right)$.
3. Deduce a formula involving ϕ and σ^2 indicating whether X Granger-causes Y.
4. Does X instantaneously Granger-cause Y?
5. Express $\mathrm{proj}\left(Y_t | \mathcal{H}^Y_{t-1}\right)$ using Y_{t-1} and ζ_{t-1} and then X_{t-1}, ϵ and η.
6. Does X Granger-cause Y ?
7. Define $\tilde{X} = \mathrm{S}\, X$. Show that $\tilde{X}$ does not instantaneously Granger-cause Y. What is the Granger causality lag in this case ?

Exercise 6.10. Consider the same setting as in Theorem 6.8.4, and define the structural form representation as in (6.20). For any vector $\mathbf{x} \in \mathbb{C}^d$, we denote by $\mathbf{x}(-d)$ the $\mathbb{C}^{d-1}$ vector obtained from the $d-1$ first entries of $\mathbf{x}$,
$$\mathbf{x}(-d) = \begin{bmatrix} \mathbf{x}(1) \\ \vdots \\ \mathbf{x}(d-1) \end{bmatrix} .$$
Similarly if A is a matrix with d columns, $A(\cdot, d)$ and $A(\cdot, -d)$ denote the last column of A and the matrix obtained from the $d-1$ first columns, respectively.

1. Show that, for all $t \geq 1$ and $j = 1, \ldots, d-1$,
$$\mathrm{proj}\left(\mathbf{X}_t(j) | \mathcal{H}^{\mathbf{X}}_{t-1}\right) = \Phi_1(j,d)\mathbf{X}_{t-1}(d) + \Phi_1(j,-d)\mathbf{X}_{t-1}(-d) + \sum_{k=2}^{p} \left[\Phi_k \mathbf{X}_{t-k}\right](j) + \sum_{k=1}^{q} \left[\Theta_k \mathbf{Z}_{t-k}\right](j) \ .$$

2. Show that, for all $t \geq 1$,

$$\mathbf{X}_t(d) - \text{proj}\left(\mathbf{X}_t(d) | \, \mathcal{H}_{t-1}^{\mathbf{X}} + \text{Span}\left(\mathbf{X}_t(-d)\right)\right) = \tilde{\mathbf{Z}}_t(d) \; .$$

We temporarily assume $q = 0$.

3. Deduce that, for all $t \geq 2$,

$$\text{proj}\left(\mathbf{X}_t(j) | \, \mathcal{H}_{t-1}^{\mathbf{X}}\right) - \text{proj}\left(\mathbf{X}_t(j) | \, \mathcal{H}_{t-1}^{\mathbf{X}(-d)} + \mathcal{H}_{t-2}^{\mathbf{X}(d)}\right) = \Phi_1(j,d)\tilde{\mathbf{Z}}_{t-1}(d) \; .$$

4. Give a necessary and sufficient condition for $\mathbf{X}(d)$ to have $\mathbf{X}(-d)$-Granger causality lag equal to 1 to predict $\mathbf{X}(j)$ for $j = 1, \ldots, d-1$.

We now assume $q \geq 1$ with Σ diagonal (hence L in (6.19) is the identity matrix).

5. Show that, for $t \geq 1$, $\mathcal{H}_{t-1}^{\mathbf{X}} + \text{Span}\left(\mathbf{X}_t(\ell)\right) = \mathcal{H}_{t-1}^{\mathbf{Z}} + \text{Span}\left(\mathbf{Z}_t(\ell)\right)$ for all $\ell = 1, \ldots, d$ and give

$$\mathbf{Z}_t(\ell) - \text{proj}\left(\mathbf{Z}_t(\ell) | \, \mathcal{H}_{t-1}^{\mathbf{X}} + \text{Span}\left(\mathbf{X}_t(-d)\right)\right)$$

6. Deduce that, for all $t \geq 2$,

$$\text{proj}\left(\mathbf{X}_t(j) | \, \mathcal{H}_{t-1}^{\mathbf{X}}\right) - \text{proj}\left(\mathbf{X}_t(j) | \, \mathcal{H}_{t-1}^{\mathbf{X}(-d)} + \mathcal{H}_{t-2}^{\mathbf{X}(d)}\right) = \left(\Phi_1(j,d) + \Theta_1(j,d)\right)\tilde{\mathbf{Z}}_{t-1}(d) \; .$$

7. Give a necessary and sufficient condition for $\mathbf{X}(d)$ to have $\mathbf{X}(-d)$-Granger causality lag equal to 1 to predict $\mathbf{X}(j)$ for $j = 1, \ldots, d-1$.

8. Consider the processes $\mathbf{X} = (X, Y)$ and $\mathbf{X} = (\tilde{X}, Y)$, successively with X, Y and $\tilde{X}$ as in Exercise 6.9. Answer Questions 4, 6 7 using the VARMA representations for these $\mathbf{X}$'s.

Chapter 7

Dynamic linear models

In this chapter, we introduce a very general and widespread approach for modeling time series: the *state-space* model. More precisely we will focus in this chapter on the *linear* state space model or *dynamic linear model* (DLM). A quite interesting feature of this class of models is the existence of efficient algorithms for forecasting or *filtering*. The latter consist in the estimation of a *hidden* variable involved in the model description.

7.1 Dynamic linear models (DLM)

Let us introduce a very general approach for modelling time series: the *state-space* models. Such an approach was first used in Kalman [1960], Kalman and Bucy [1961] for space tracking, where the state equation models the motion of the position of a spacecraft with location $\mathbf{X}_t$ and the data $\mathbf{Y}_t$ represents the information that can be observed from a tracking device such as velocity and azimuth. Here we focus on the *linear* state space model.

Definition 7.1.1 (DLM)**.** *A multivariate process $(\mathbf{Y}_t)_{t\geq 1}$ is said to be the* observation variables *of a* linear state-space model *or* DLM *if there exists a process $(\mathbf{X}_t)_{t\geq 1}$ of* state variables *such that that Assumption 7.1.1 below holds. The space of the state variables $\mathbf{X}_t$ (here $\mathbb{R}^p$ or $\mathbb{C}^p$) is called the* state space *and the space of the observation variables $\mathbf{Y}_t$ (here $\mathbb{R}^q$ or $\mathbb{C}^q$) is called the* observation space*.*

Assumption 7.1.1. *$(\mathbf{X}_t)_{t\geq 1}$ and $(\mathbf{Y}_t)_{t\geq 1}$ are p-dimensional and q-dimensional time series satisfying the following equations for all $t \geq 1$,*

$$\mathbf{X}_t = \Phi_t \mathbf{X}_{t-1} + \mathrm{A}_t \mathbf{u}_t + \mathbf{W}_t \,, \tag{7.1}$$

$$\mathbf{Y}_t = \Psi_t \mathbf{X}_t + \mathrm{B}_t \mathbf{u}_t + \mathbf{V}_t \,, \tag{7.2}$$

where

(i) $(\mathbf{W}_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, Q)$ *where Q is a $p \times p$ covariance matrix.*

(ii) $(\mathbf{u}_t)_{t\in\mathbb{N}}$ *is an r-dimensional exogenous input series and A_t a $p \times r$ matrix of parameters, which is possibly the zero matrix.*

(iii) The initial state $\mathbf{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_0)$.

(iv) Ψ_t *is a $q \times p$* measurement *or* observation matrix *for all $t \geq 1$,*

(v) The matrix B_t *is a* $q \times r$ *regression matrix which may be the zero matrix.*

(vi) $(\mathbf{V}_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, R)$ *where* R *is a* $q \times q$ *covariance matrix.*

(vii) The initial state $\mathbf{X}_0$*, the state noise* $(\mathbf{W}_t)_{t\geq 1}$ *and the observation noise* $(\mathbf{V}_t)_{t\geq 1}$ *are independent.*

The Gaussian Assumption will be heavily used in particular through the computation of conditional expectations. By Proposition B.4.8, if $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ is determined by the L^2 projection

$$\operatorname{proj}\left(\mathbf{X}|\left\{a + B\mathbf{Y} \,:\, a \in \mathbb{R}^p\,,\; B \in \mathbb{R}^{p\times q}\right\}\right)\,, \tag{7.3}$$

and by the covariance matrix of the error. Conversely, an important consequence of this proposition is that many computations done in this chapter continue to hold when the Gaussian assumption is dropped, provided that conditional expectations of the form $\mathbb{E}\left[\mathbf{X}|\,\mathbf{Y}\right]$ are replaced by (7.3), see Corollary 7.2.3.

Remark 7.1.1. *A slight extension of this model is to let the covariance matrices* R *and* Q *depend on* t*. All the results of Section Section 7.2 are carried out in the same way in this situation. Nevertheless, we do not detail this case here for sake of simplicity.*

The *state equation* (7.1) determines how the $p \times 1$ state vector $\mathbf{X}_t$ is generated from the past $p \times 1$ state $\mathbf{X}_{t-1}$. The *observation equation* (7.2) describes how the observed data is generated from the state data.

As previously mentioned, the model is quite general and can be used in a number of problems from a broad class of disciplines. We will see a few examples in this chapter.

Example 7.1.1 (Noisy observations of a random trend)**.** *Let us first use the state space model to simulate an artificial time series. Let* $\beta \in \mathbb{R}$*,* Z_1 *be a Gaussian random variable and* (W_t) *be a Gaussian white noise* $\mathrm{IID}(0, \sigma^2)$ *uncorrelated with* Z_1 *and define, for all* $t \geq 1$*,*

$$Z_{t+1} = Z_t + \beta + W_t = Z_1 + \beta t + W_1 + \cdots + W_t\,, t \geq 0\,.$$

When σ *is low,* Z_t *is approximatively linear with respect to* t*. The noise* (W_t) *introduce a random fluctuation around this linear trend. A noisy observation of* (Z_t) *is defined as*

$$Y_t = Z_t + V_t\,,$$

where (V_t) *is a Gaussian white noise uncorrelated with* (W_t) *and* Z_1*.*

A state-space representation of (Z_t) *can be defined by setting* $X_t = [Z_t, \beta]^T$*, so that the state equation reads*

$$X_{t+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} X_t + V_t \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,.$$

The observation equation is then $Y_t = [1\ \ 0]X_t + V_t$ *The process* (Z_t) *is obtained from* (X_t) *by* $Z_t = [1\ \ 0]X_t$*. We display a simulated* (Z_t) *and* (Y_t) *in Figure Figure 7.1.*

Example 7.1.2 (Climatology data)**.** *Figure 7.2 shows two different estimates of the global temperature deviations from 1880 to 2009. They can be found on the site*

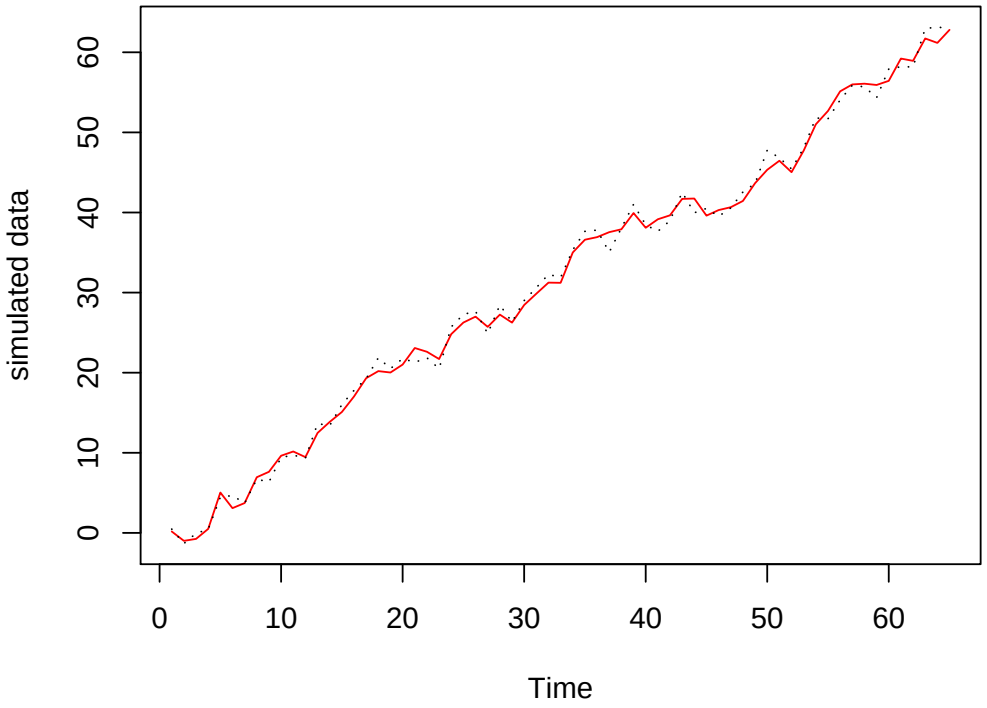

Figure 7.1: Simulated random trend (plain red line) and its observation with additive noise (dotted black line).

http://data.giss.nasa.gov/gistemp/graphs/.

The solid red line represents the global mean land-ocean temperature index data. The dotted black line represents the surface-air temperature index data using only land based meteorological station data. Thus, both series are measuring the same underlying climate signal but with different measurement conditions. From a modelling point of view, we may suggest the following observation equations

$$Y_{1,t} = X_t + V_{1,t} \quad \text{and} \quad Y_{2,t} = X_t + V_{2,t},$$

or more compactly,

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} X_t + \begin{bmatrix} V_{1,t} \\ V_{2,t} \end{bmatrix},$$

where

$$R = \text{Cov} \begin{bmatrix} V_{1,t} \\ V_{2,t} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}.$$

The unknown common signal X_t also needs some evolution equation. A natural one is the random walk with drift which states

$$X_t = \delta + X_{t-1} + W_t,$$

where $(W_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, Q)$. *In this example,* $p = 1$, $q = 2$, $\Phi_t = 1$, $\mathrm{A}_t = \delta$ *with* $\mathbf{u}_t = 1$, *and* $\mathrm{B}_t = 0$.

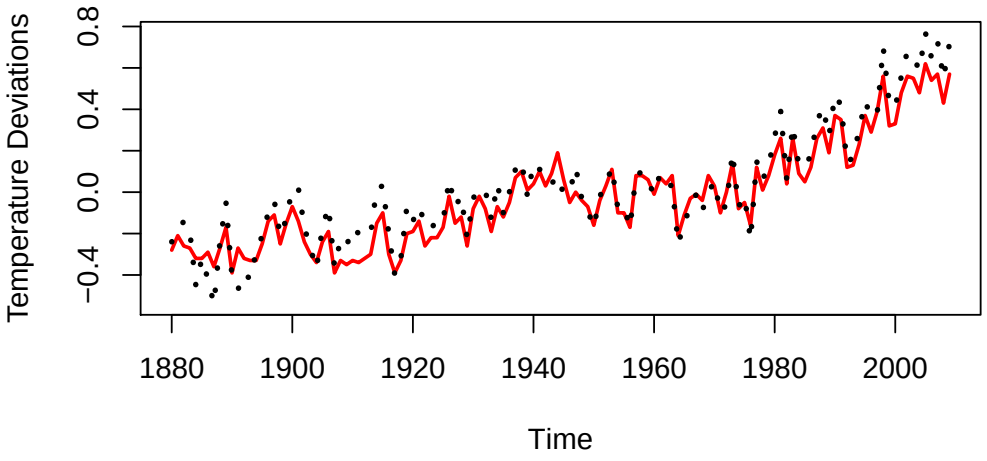

Figure 7.2: Annual global temperature deviation series, measured in degrees centigrade, 1880–2009.

Dynamic linear models allow us to provide a quite general framework for denoising and forecasting a Gaussian process, or/and estimating its parameters. In (7.1) and (7.2), unknown parameters are possibly contained in $\Phi_t, \mathrm{A}_t, Q, \mathrm{B}_t, \Psi_t$, and R that define the particular model. It is also of interest to estimate (or *denoise*) and to forecast values of the underlying unobserved process $(\mathbf{X}_t)_{t\in\mathbb{N}}$. It is important to mention that a large family of stationary Gaussian processes enter this general framework, as shown in the last following simple example.

Example 7.1.3 (Noisy AR(1) process)**.** *Consider a stationary process satisfying the AR(1) equation*

$$X_t = \phi X_{t-1} + W_t, \quad t \in \mathbb{Z}\,,$$

where $|\phi| < 1$ *and* $(W_t)_{t\in\mathbb{Z}} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_w^2)$. *Then using the results of Section 3.3, we easily get that the autocovariance function of* $(X_t)_{t\in\mathbb{N}}$ *is*

$$\gamma_x(h) = \frac{\sigma_w^2}{1-\phi^2}\ \phi^{|h|}, \quad h = 0, \pm 1, \pm 2, \dots\ ,$$

and $X_0 \sim \mathcal{N}\left(0, \sigma_w^2/(1-\phi^2)\right)$ *is independent of* $(W_t)_{t\in\mathbb{N}}$. *Suppose now that we observe a noisy version of* $(X_t)_{t\in\mathbb{N}}$, *namely*

$$Y_t = X_t + V_t,$$

where $(V_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2)$ *and* $(V_t)_{t\in\mathbb{N}}$ *and* $(W_t)_{t\in\mathbb{Z}}$ *are independent. Then the observations are stationary because* $(Y_t)_{t\in\mathbb{N}}$ *is the sum of two independent stationary components* $(X_t)_{t\in\mathbb{N}}$

and $(V_t)_{t\in\mathbb{N}}$. *Simulated series* X_t *and* Y_t *with* $\phi = 0.8$ *and* $\sigma_w = \sigma_v = 1.0$ *are displayed in Figure 7.3. We easily compute*

$$\gamma_y(0) = \mathrm{Var}\,(Y_t) = \mathrm{Var}\,(X_t + V_t) = \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2, \tag{7.4}$$

and, when $h \neq 0$,

$$\gamma_y(h) = \mathrm{Cov}\,(Y_t, Y_{t-h}) = \mathrm{Cov}\,(X_t + V_t, X_{t-h} + V_{t-h}) = \gamma_x(h).$$

Consequently, for $h \neq 0$, *the ACF of the observations is*

$$\rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_v^2}{\sigma_w^2}(1-\phi^2)\right)^{-1} \phi^{|h|}.$$

It can be shown that $(Y_t)_{t\in\mathbb{Z}}$ *is an ARMA(1,1) process (see Exercise 7.1). We will provide a general view on the relationships between DLMs and stationary ARMA processes in Section 7.5.*

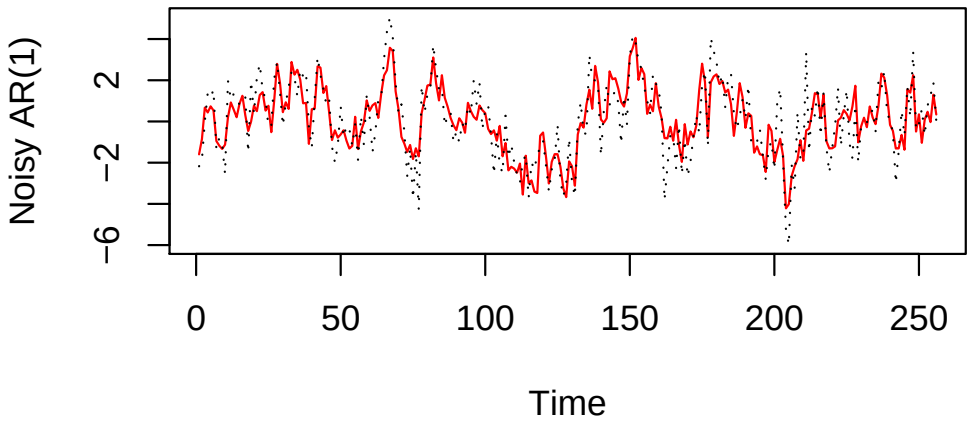

Figure 7.3: Simulated AR(1) process (solid red) and a noisy observation of it (dotted black).

7.2 Kalman approach for filtering, forecasting and smoothing

The state-space models are primarily used for estimating the underlying unobserved signal $\mathbf{X}_t$, given the data $\mathbf{Y}_{1:s} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_s\}$. More precisely, it consists in computing the conditional mean

$$\mathbf{X}_{t|s} \overset{\text{def}}{=} \mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{Y}_{1:s}\right] \tag{7.5}$$

and to measure the L^2 norm of the error $\mathbf{X}_t - \mathbf{X}_{t|s}$,

$$\mathbf{\Sigma}_{t|s} \stackrel{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_t - \mathbf{X}_{t|s})(\mathbf{X}_t - \mathbf{X}_{t|s})^T\right] = \text{Cov}(\mathbf{X}_t - \mathbf{X}_{t|s}) \,, \tag{7.6}$$

since $\mathbf{X}_t - \mathbf{X}_{t|s}$ is centered.

Three different situations are generally distinguished.

a- It is called a *forecasting* or prediction problem if $s < t$.

b- It is called a *filtering* problem if $s = t$.

c- It is called a *smoothing* problem if $s > t$.

Interestingly, these tasks are very much related to the computation of the likelihood for estimating the unknown parameters of the models, see Section 7.6.

The Kalman filter is a recursive algorithm that provides an efficient way to compute the filtering and first order forecasting equations $\mathbf{X}_{t|t-1}$ and $\mathbf{X}_{t|t}$. It is defined as follows.

Algorithm 6: Kalman filter algorithm.

Data: Parameters Q, R and A_t, B_t, Ψ_t for $t = 1, \ldots, n$, initial conditions $\boldsymbol{\mu}$ and Σ_0, observations $\mathbf{Y}_t$ and exogenous input series $\mathbf{u}_t$, for $t = 1, \ldots, n$.

Result: Forecasting and filtering outputs $\mathbf{X}_{t|t-1}$, $\mathbf{X}_{t|t}$, and their autocovariance matrices $\mathbf{\Sigma}_{t|t-1}$ and $\mathbf{\Sigma}_{t|t}$ for $t = 1, \ldots, n$.

Initialization: set $\mathbf{X}_{0|0} = \boldsymbol{\mu}$ and $\mathbf{\Sigma}_{0|0} = \Sigma_0$.

for $t = 1, 2, \ldots, n$ **do**

Compute in this order

$$\mathbf{X}_{t|t-1} = \Phi_t \mathbf{X}_{t-1|t-1} + \mathrm{A}_t \mathbf{u}_t, \tag{7.7}$$

$$\mathbf{\Sigma}_{t|t-1} = \Phi_t \mathbf{\Sigma}_{t-1|t-1} \Phi_t^T + Q, \tag{7.8}$$

$$K_t = \mathbf{\Sigma}_{t|t-1} \Psi_t^T [\Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R]^{-1}. \tag{7.9}$$

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + K_t(\mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t), \tag{7.10}$$

$$\mathbf{\Sigma}_{t|t} = [I - K_t \Psi_t] \mathbf{\Sigma}_{t|t-1}. \tag{7.11}$$

end

Proposition 7.2.1 (Kalman Filter)**.** *Algorithm 6 holds for the state-space model satisfying Assumption 7.1.1, provided that $\Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R$ are invertible matrices for $t = 1, \ldots, n$.*

The matrix K_t defined in (7.9) is called the *Kalman gain matrix*.

For proving Proposition 7.2.1, we will introduce the following definition

$$\begin{aligned} \mathbf{Y}_{t|s} &\stackrel{\text{def}}{=} \mathbb{E}\left[\mathbf{Y}_t \middle| \mathbf{Y}_{1:s}\right] \\ \boldsymbol{\epsilon}_t &\stackrel{\text{def}}{=} \mathbf{Y}_t - \mathbf{Y}_{t|t-1} \\ \Gamma_t &\stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T\right] = \text{Cov}(\boldsymbol{\epsilon}_t) \,, \end{aligned}$$

and show the folowing useful formula

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t, \tag{7.12}$$

$$\Gamma_t = \text{Cov}(\Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t) = \Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R \tag{7.13}$$

for $t = 1, \ldots, n$. The process $(\boldsymbol{\epsilon}_t)$ is called the *innovation process* of $(\mathbf{Y}_t)$.

Proof of Proposition 7.2.1. By Assumption 7.1.1, we have that $(\mathbf{W}_t)_{t>s}$ is independent of $\mathbf{Y}_{1:s}$ and $\mathbf{X}_{1:s}$ and $(\mathbf{V}_t)_{t>s}$ is independent of $\mathbf{Y}_{1:s}$ and $(\mathbf{X}_t)_{t\geq 0}$.

Using (7.1), this implies that, for all $t > s$

$$\begin{aligned}\mathbf{X}_{t|s} &= \mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{Y}_{1:s}\right] \\ &= \mathbb{E}\left[\Phi_t \mathbf{X}_{t-1} + \mathrm{A}_t \mathbf{u}_t + \mathbf{W}_t \middle| \mathbf{Y}_{1:s}\right] \\ &= \Phi_t \mathbf{X}_{t-1|s} + \mathrm{A}_t \mathbf{u}_t \ , \end{aligned} \tag{7.14}$$

and, moreover,

$$\begin{aligned}\mathbf{\Sigma}_{t|s} &= \mathrm{Cov}\left(\mathbf{X}_t - \mathbf{X}_{t|s}\right) \\ &= \mathrm{Cov}\left(\Phi_t(\mathbf{X}_{t-1} - \mathbf{X}_{t-1|s}) + \mathbf{W}_t\right) \\ &= \Phi_t \mathbf{\Sigma}_{t-1|s} \Phi_t^T + Q \ . \end{aligned} \tag{7.15}$$

which gives (7.7) and (7.8).

Next, we show (7.10). By definition of the innovation process, the σ-field generated by $\mathbf{Y}_{1:t}$ is the same as that generated by $\mathbf{Y}_{1:t-1}$ and $\boldsymbol{\epsilon}_t$, thus we have

$$\mathbf{X}_{t|t} = \mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{Y}_{1:t-1}, \boldsymbol{\epsilon}_t\right] \ .$$

By Assumption 7.1.1, the variables $\mathbf{X}_t$, $\mathbf{Y}_{1:t-1}$ and $\boldsymbol{\epsilon}_t$ are jointly Gaussian. It then follows from Proposition Proposition B.4.8 that

$$\mathbf{X}_{t|t} = \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(\left\{a,\ B\mathbf{Y}_s,\ B\boldsymbol{\epsilon}_t\ :\ a \in \mathbb{R}^p\,,\ B \in \mathbb{R}^{p\times q},\ 1 \leq s \leq t-1\right\}\right)\right) \ ,$$

Observing that $\boldsymbol{\epsilon}_t$ is centered and uncorrelated with $\mathbf{Y}_{1:t-1}$, we get that

$$\begin{aligned}\mathbf{X}_{t|t} = \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(\left\{a,\ B\mathbf{Y}_s\ :\ a \in \mathbb{R}^p\,,\ B \in \mathbb{R}^{p\times q},\ 1 \leq s \leq t-1\right\}\right)\right) \\ + \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(\left\{B\boldsymbol{\epsilon}_t\ :\ B \in \mathbb{R}^{p\times q}\right\}\right)\right) \ ,\end{aligned}$$

and thus, setting

$$K_t = \mathrm{Cov}\left(\mathbf{X}_t, \boldsymbol{\epsilon}_t\right)\mathrm{Cov}(\boldsymbol{\epsilon}_t)^{-1} = \mathrm{Cov}\left(\mathbf{X}_t, \mathbf{Y}_t - \mathbf{Y}_{t|t-1}\right)\Gamma_t^{-1} \ ,$$

we have

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + K_t\boldsymbol{\epsilon}_t \ ,$$

and

$$\begin{aligned}\mathbf{\Sigma}_{t|t} &= \mathbf{\Sigma}_{t|t-1} - \mathrm{Cov}(K_t\boldsymbol{\epsilon}_t) \\ &= \mathbf{\Sigma}_{t|t-1} - K_t\Gamma_t K_t^T \\ &= \mathbf{\Sigma}_{t|t-1} - K_t\mathrm{Cov}\left(\mathbf{X}_t, \mathbf{Y}_t - \mathbf{Y}_{t|t-1}\right)^T \ .\end{aligned}$$

Now, by (7.2), we have

$$\mathbf{Y}_{t|t-1} = \mathbb{E}\left[\Psi_t\mathbf{X}_t + \mathrm{B}_t\mathbf{u}_t + \mathbf{V}_t \middle| \mathbf{Y}_{1:t-1}\right] = \Psi_t\mathbf{X}_{t|t-1} + \mathrm{B}_t\mathbf{u}_t \ ,$$

and thus

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{X}_t, \mathbf{Y}_t - \mathbf{Y}_{t|t-1}\right) &= \mathrm{Cov}\left(\mathbf{X}_t, \Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t\right) \\ &= \mathbf{\Sigma}_{t|t-1}\Psi_t^T \ ,\end{aligned}$$

and

$$\begin{aligned}\Gamma_t &= \mathrm{Cov}(\mathbf{Y}_t - \mathbf{Y}_{t|t-1}) \\ &= \mathrm{Cov}\left(\Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t\right) \\ &= \Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R \,.\end{aligned}$$

Hence, we finally get that

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + K_t(\mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t) \,,$$

and

$$\mathbf{\Sigma}_{t|t} = \mathbf{\Sigma}_{t|t-1} - K_t \Psi_t \mathbf{\Sigma}_{t|t-1} \,,$$

with

$$K_t = \mathbf{\Sigma}_{t|t-1} \Psi_t^T [\Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R]^{-1} \,.$$

That is, we have shown (7.9), (7.10) and (7.11) and the proof is concluded. □

Let us consider the forecasting and smoothing problems, that is the computation of $\mathbf{X}_{t|n}$ for $t > n$ and $t = 1, \ldots, n-1$, successively. These algorithms complete Algorithm 6 in the sense that in practice one can use them after having first applied Algorithm 6.

Algorithm 7: Kalman forecasting algorithm.

Data: A forecasting lag h, parameters Q and A_t for $t = n+1, \ldots, n+h$, and exogenous input series $\mathbf{u}_t$, for $t = n+1, \ldots, n+h$, Kalman filter output $\mathbf{X}_{n|n}$ and its error matrix $\mathbf{\Sigma}_{n|n}$.

Result: Forecasting output $\mathbf{X}_{t|n}$ and their error matrices $\mathbf{\Sigma}_{t|n}$ for $t = n+1, \ldots, n+h$

Initialization: set $k = 1$.

for $k = 1, 2, \ldots, h$ **do**

Compute in this order

$$\begin{aligned}\mathbf{X}_{n+k|n} &= \Phi_{n+k} \mathbf{X}_{n+k-1|n} + \mathrm{A}_{k+n} \mathbf{u}_{n+k} \,, \\ \mathbf{\Sigma}_{n+k|n} &= \Phi_{n+k} \mathbf{\Sigma}_{n+k-1|n} \Phi_{n+k}^T + Q \,.\end{aligned}$$

end

Algorithm 8: Rauch-Tung-Striebel smoother algorithm.

Data: Parameters Φ_t for $t = 1, \ldots, n$, and exogenous input series $\mathbf{u}_t$, for $t = n+1, \ldots, n+h$, Kalman filter output $\mathbf{X}_{t|t}$, $\mathbf{X}_{t|t-1}$, and their error matrices $\mathbf{\Sigma}_{t|t}$ and $\mathbf{\Sigma}_{t|t-1}$ for $t = 1, \ldots, n$.

Result: Smoothing outputs $\mathbf{X}_{t|n}$, and their autocovariance matrices $\mathbf{\Sigma}_{t|n}$ for $t = n-1, n-2 \ldots, 1$.

for $t = n, n-1, \ldots, 2$ **do**

Compute in this order

$$J_{t-1} = \mathbf{\Sigma}_{t-1|t-1} \Phi_t^T \mathbf{\Sigma}_{t|t-1}^{-1} \,, \tag{7.16}$$

$$\mathbf{X}_{t-1|n} = \mathbf{X}_{t-1|t-1} + J_{t-1}\left(\mathbf{X}_{t|n} - \mathbf{X}_{t|t-1}\right), \tag{7.17}$$

$$\mathbf{\Sigma}_{t-1|n} = \mathbf{\Sigma}_{t-1|t-1} + J_{t-1}\left(\mathbf{\Sigma}_{t|n} - \mathbf{\Sigma}_{t|t-1}\right) J_{t-1}^T \,. \tag{7.18}$$

end

Proposition 7.2.2. *Algorithm 7 and Algorithm 8 hold for the state-space model satisfying Assumption 7.1.1, provided that (only for Algorithm 8)* $\mathbf{\Sigma}_{t|t-1}$ *is an invertible matrix for* $t = 2, \dots, n$.

Proof. Algorithm 7 directly follows from (7.14) and (7.15).

We now show that Algorithm 8 holds. Observe that $\mathbf{Y}_{1:n}$ can be generated with $\mathbf{Y}_{1:t-1}$, $\mathbf{X}_t$, $\mathbf{V}_{t:n}$, and $\mathbf{W}_{t+1:n}$. Thus we have

$$\mathbb{E}\left[\mathbf{X}_{t-1} \middle| \mathbf{Y}_{1:n}\right] = \mathbb{E}\left[\widetilde{\mathbf{X}}_{t-1} \middle| \mathbf{Y}_{1:n}\right] , \tag{7.19}$$

where

$$\begin{aligned}\widetilde{\mathbf{X}}_{t-1} &= \mathbb{E}\left[\mathbf{X}_{t-1} \middle| \mathbf{Y}_{1:t-1}, \mathbf{X}_t - \mathbf{X}_{t|t-1}, \mathbf{V}_{t:n}, \mathbf{W}_{t+1:n}\right] \\ &= \mathbb{E}\left[\mathbf{X}_{t-1} \middle| \mathbf{Y}_{1:t-1}, \mathbf{X}_t - \mathbf{X}_{t|t-1}\right] ,\end{aligned}$$

since $\mathbf{V}_{t:n}, \mathbf{W}_{t+1:n}$ are independent of all other variables appearing in this formula. Using the Gaussian assumption and the fact that $\mathbf{Y}_{1:t-1}$ and $\mathbf{X}_t - \mathbf{X}_{t|t-1}$ are uncorrelated, we get

$$\widetilde{\mathbf{X}}_{t-1} = \mathbf{X}_{t-1|t-1} + J_{t-1}(\mathbf{X}_t - \mathbf{X}_{t|t-1}), \tag{7.20}$$

and

$$\mathrm{Cov}\left(\mathbf{X}_{t-1} - \widetilde{\mathbf{X}}_{t-1}\right) = \mathbf{\Sigma}_{t-1|t-1} - J_{t-1}\mathbf{\Sigma}_{t|t-1}J_{t-1}^T, \tag{7.21}$$

where

$$J_{t-1} = \mathrm{Cov}(\mathbf{X}_{t-1}, \mathbf{X}_t - \mathbf{X}_{t|t-1})\mathbf{\Sigma}_{t|t-1}^{-1} = \mathbf{\Sigma}_{t-1|t-1}\Phi_t^T\mathbf{\Sigma}_{t|t-1}^{-1} ,$$

which corresponds to (7.16). By (7.19) and (7.20), and by projecting $\widetilde{\mathbf{X}}_{t-1}$ on

$$\mathrm{Span}\left(\left\{a, B\mathbf{Y}_s \,:\, a \in \mathbb{R}^p,\ B \in \mathbb{R}^{p\times q},\ 1 \le s \le n\right\}\right) ,$$

we obtain

$$\mathbf{X}_{t-1|n} = \mathbf{X}_{t-1|t-1} + J_{t-1}(\mathbf{X}_{t|n} - \mathbf{X}_{t|t-1}) ,$$

that is (7.17) and

$$\mathrm{Cov}\left(\widetilde{\mathbf{X}}_{t-1} - \mathbf{X}_{t-1|n}\right) = J_{t-1}\mathbf{\Sigma}_{t|n}J_{t-1}^T .$$

This, with (7.21), and using that $\widetilde{\mathbf{X}}_{t-1} - \mathbf{X}_{t-1|n}$ and $\mathbf{X}_{t-1} - \widetilde{\mathbf{X}}_{t-1}$ are uncorrelated, we obtain

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{X}_{t-1} - \mathbf{X}_{t-1|n}\right) &= \mathrm{Cov}\left(\mathbf{X}_{t-1} - \widetilde{\mathbf{X}}_{t-1} + \widetilde{\mathbf{X}}_{t-1} - \mathbf{X}_{t-1|n}\right) \\ &= \mathbf{\Sigma}_{t-1|t-1} - J_{t-1}\mathbf{\Sigma}_{t|t-1}J_{t-1}^T + J_{t-1}\mathbf{\Sigma}_{t|n}J_{t-1}^T ,\end{aligned}$$

that is (7.18). □

Inspecting the proofs of Proposition 7.2.1 and Proposition 7.2.2, we have the following result which says that if the Gaussian assumption is dropped, then the above algorithms continues to hold in the framework of linear prediction.

Corollary 7.2.3. *Suppose that Assumption 7.1.1 holds but with* $\mathbf{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_0)$, $(\mathbf{V}_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, R)$ *and* $(\mathbf{W}_t)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}(0, Q)$ *replaced by the weaker conditions* $\mathbb{E}[\mathbf{X}_0] = \boldsymbol{\mu}$, $\mathrm{Cov}(\mathbf{X}_0) = \Sigma_0$, $(\mathbf{V}_t)_{t\in\mathbb{N}} \sim \mathrm{WN}(0, R)$, $(\mathbf{W}_t)_{t\in\mathbb{N}} \sim \mathrm{WN}(0, Q)$. *Then Algorithm 6, Algorithm 7 and Algorithm 8 continue to hold if the definitions of* $\mathbf{X}_{s|t}$ *in (7.7) is replaced by*

$$\mathbf{X}_{s|t} \overset{\text{def}}{=} \mathrm{proj}\left(\mathbf{X}_s \middle| \mathrm{Span}\left(\left\{a, B\mathbf{Y}_s \,:\, a \in \mathbb{R}^p,\ B \in \mathbb{R}^{p\times q},\ 1 \le s \le t\right\}\right)\right) .$$

For estimation purposes, we will need to compute the one-lag covariance matrix of the smoother outputs that is

$$\boldsymbol{\Sigma}_{t_1,t_2|s} \overset{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_{t_1} - \mathbf{X}_{t_1|s})(\mathbf{X}_{t_2} - \mathbf{X}_{t_2|s})^T\right] \tag{7.22}$$

with $t_1 = t, t_2 = t-1$ and $s = n$. Note that this notation extends the previous one in the sense that $\boldsymbol{\Sigma}_{t|s} = \boldsymbol{\Sigma}_{t,t|s}$.

One simple way to compute $\boldsymbol{\Sigma}_{t-1,t-2|n}$ is to define new state and observation variables by stacking two consecutive times together, namely

$$\mathbf{X}_{(t)} \overset{\text{def}}{=} [\mathbf{X}_t^T\ \mathbf{X}_{t-1}^T]^T ,$$
$$\mathbf{Y}_{(t)} \overset{\text{def}}{=} [\mathbf{Y}_t^T\ \mathbf{Y}_{t-1}^T]^T .$$

Here the parentheses around the time variable t indicate that we are dealing with the stacked variables. One can deduce the state and observation equations for these variables and apply the Kalman filter and smoother to compute

$$\boldsymbol{\Sigma}_{(t)|(n)} = \begin{bmatrix} \boldsymbol{\Sigma}_{t|n} & \boldsymbol{\Sigma}_{t|t-1n} \\ \boldsymbol{\Sigma}_{t,t-1|n}^T & \boldsymbol{\Sigma}_{t-1|n} \end{bmatrix},$$

where subscripts (t) and (n) again refer to operations on the stacked values.

However there is a more direct and more convenient way to compute these covariances. The proof of validity of the following algorithm is left to the reader (see Exercise 7.2).

Algorithm 9: One-lag covariance algorithm.

Data: Parameters Ψ_n and Φ_t for $t = 1, \ldots, n$, Gain matrix K_n Kalman filter covariance matrices $\boldsymbol{\Sigma}_{t|t}$ and $\boldsymbol{\Sigma}_{t|t-1}$ for $t = 1, \ldots, n$, matrices J_t for $t = 1, \ldots, n-1$.

Result: One-lag covariance matrices for smoother outputs $\boldsymbol{\Sigma}_{t,t-1|n}$ for $t = 1, \ldots, n$.

Initialization: Set

$$\boldsymbol{\Sigma}_{n,n-1|n} = (I - K_n\Psi_n)\Phi_n\boldsymbol{\Sigma}_{n-1|n-1}, \tag{7.23}$$

for $t = n, n-1, \ldots, 2$ **do**

$$\boldsymbol{\Sigma}_{t-1,t-2|n} = \boldsymbol{\Sigma}_{t-1|t-1}J_{t-2}^T + J_{t-1}\left(\boldsymbol{\Sigma}_{t,t-1|n} - \Phi_t\boldsymbol{\Sigma}_{t-1|t-1}\right)J_{t-2}^T. \tag{7.24}$$

end

Remark 7.2.1. *All the above algorithms (Algorithms 6, 7, 8 and 9) are* recursive *in the sense that their outputs are computed using a simple recursive set of equations. Algorithm 6 can moreover be implemented* online *in the sense that each iteration of the recursion at time* t *only uses* ***one new observation*** $\mathbf{Y}_t$***, without having to reprocess the entire data set*** $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$. *Since the number of computations at each iteration is constant (*$O(1)$ *operations at each step), it means that in practice, it can be run at the same time as the acquisition of the observed data.*

Remark 7.2.2. *It is interesting to note that in the above algorithms, the covariance matrices do not depend on the observations* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$*, only on the parameters of the dynamic linear model. Hence, if these parameters are known (as assumed in this section), they can be computed* off-line*, in particular before having acquired the observations* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$*.*

Example 7.2.1 (Noisy AR(1) (continued from Example 7.1.3))**.** *Let* $(X_t)_{t\in\mathbb{Z}}$ *and* $(Y_t)_{t\in\mathbb{Z}}$ *be as in Example 7.1.3. We apply the algorithms using the* true *parameters used for generating the data, namely,* $\mathrm{A}_t = 0$*,* $\Phi = \phi$*,* $Q = \sigma_w^2$*,* $\Psi = 1$*,* $\mathrm{B} = 0$*,* $R = \sigma_v^2$ $\mu_0 = 0$ *and* $\Sigma_0 = \gamma_y(0) = \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2$ *(see (7.4)). To produce Figure 7.4, the Kalman smoother was computed with these true parameters from* $Y_1, \ldots, Y_n$ *with* $n = 2^8$ *only the last 16 points of* Y_t*,* X_t *and* $\mathbf{X}_{t|n}$ $(t = n-15, n-14, \ldots, n)$ *are drawn, using respectively red circles, a dotted black line and a solid green line. The dashed blue lines represent 95% confidence intervals for* $\mathbf{X}_t$ *obtained using that, given* $\mathbf{Y}_{1:n}$*, the conditional distribution of each* $\mathbf{X}_t$ *is* $\mathcal{N}(\mathbf{X}_{t|n}, \mathbf{\Sigma}_{t|n})$*.*

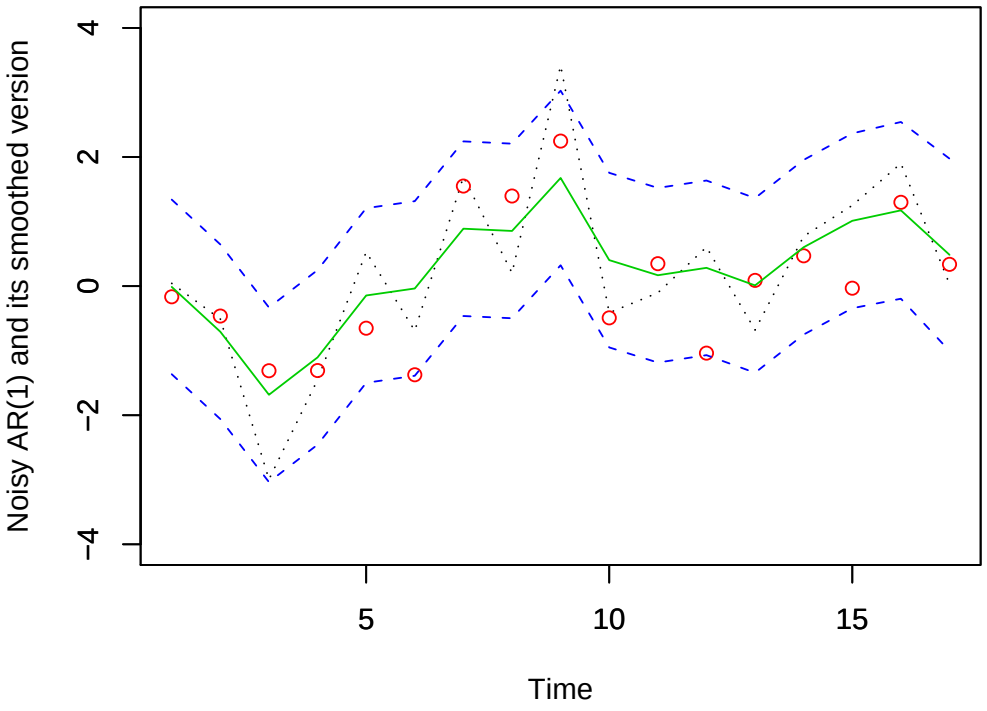

Figure 7.4: Simulated AR(1) process (red circles), a noisy observation of it (dotted black line), the smoother outputs (solid green line) and the 95% confidence intervals (between blued dashed lines).

7.3 Steady State approximations

Let us consider Assumption 7.1.1 in the particular case where there are no input series ($\mathrm{A}_t = \mathrm{B}_t = 0$) and the observation and state equation does not vary along the time ($\Phi_t = \Phi$ and $\Psi_t = \Psi$). If moreover the state equations yields a time series $(\mathbf{X}_t)$ which "looks" stationary, then one can expect that the distribution of $(\mathbf{X}_{1:n}, \mathbf{Y}_{1:n})$ yields steady equations for filtering, that is, in Algorithm 6, the Kalman gain K_t and the error covariance matrices $\mathbf{\Sigma}_{t|t}$ and $\mathbf{\Sigma}_{t|t-1}$ should not depend on t. Of course, this cannot be exactly true : these quantities correspond to state and observation variables $(\mathbf{X}_{1:t}, \mathbf{Y}_{1:t})$ whose distribution cannot be exactly the same as $((\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1})$. But it can be approximately true if the past data has a very small influence on the current ones, in other words, if the conditional distribution of $\mathbf{X}_t$ given $\mathbf{Y}_{1:t}$ is approximately the same as the conditional distribution of $\mathbf{X}_t$ given the whole past $\mathbf{Y}_{-\infty:t}$.

In practice this steady approximation of the Kalman filter is observed when $K_t \to K$ and $\mathbf{\Sigma}_{t|t-1} \to \Sigma$ as $t \to \infty$. Using the relationship between $\mathbf{\Sigma}_{t|t-1}$ and $\mathbf{\Sigma}_{t-1|t-2}$ (following (7.11) and (7.11)), we obtain that Σ is a necessary solution to the *Ricatti equation*

$$\Sigma = \Phi[\Sigma - \Sigma\Psi^T(\Psi\Sigma\Psi^T + R)^{-1}\Psi\Sigma]\Phi^T + Q \;, \tag{7.25}$$

and following (7.9), the steady-state gain matrix reads

$$K = \Sigma\Psi^T[\Psi\Sigma\Psi^T + R]^{-1} \;.$$

The convergence of the MLE and its asymptotic normality, stated in (4.37), can be established when Φ has eigenvalues within the open unit disk $\{z \in \mathbb{C}\,,\, |z| < 1\}$. We just refer to Caines [1988], Hannan and Deistler [1988] for details. Let us just briefly give a hint of why this assumption is meaningful. Iterating the state equation (7.1) in the case $\Phi_t = \Phi$ and $\mathrm{A}_t = 0$ yields

$$\mathbf{X}_t = \Phi^t\mathbf{X}_0 + \sum_{k=0}^{t-1} \Phi^k \boldsymbol{\epsilon}_{t-k} \;.$$

Thus, if the spectral radius of Φ is strictly less than 1, then $\mathbf{X}_t$ can be approximated by the series

$$\tilde{\mathbf{X}}_t = \sum_{k=0}^{\infty} \Phi^k \boldsymbol{\epsilon}_{t-k} \;,$$

which defines a stationary process. With this stationary approximation, and using the machinery introduced in Section 4.2, one can derive the asymptotic behavior of the MLE, under appropriate assumptions of the parameterization.

7.4 Correlated Errors

Sometimes it is advantageous to use assumptions for the linear state-space model which are slightly different from Assumption 7.1.1. In the following set of assumptions, the model on the error terms $\mathbf{W}_t$ and $\mathbf{V}_t$ is modified: a matrix Θ is introduced in the state space equation and some correlation S may appear between $\mathbf{V}_t$ and $\mathbf{W}_t$. We say that the linear state-space model has *correlated errors*. Note also that the indices in the state-space equation are changed so that the correlation is introduced between errors applied to the same $\mathbf{X}_t$.

Assumption 7.4.1. *Suppose that the state variables* $(\mathbf{X}_t)_{t\geq 1}$ *and the observed variables* $(\mathbf{Y}_t)_{t\geq 1}$ *are* p*-dimensional and* q*-dimensional time series satisfying the following equations for all* $t \geq 1$*,*

$$\mathbf{X}_{t+1} = \Phi_t \mathbf{X}_t + \mathrm{A}_{t+1}\mathbf{u}_{t+1} + \Theta_t \mathbf{W}_t \,, \tag{7.26}$$

$$\mathbf{Y}_t = \Psi_t \mathbf{X}_t + \mathrm{B}_t \mathbf{u}_t + \mathbf{V}_t \,, \tag{7.27}$$

where

(i) $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{V}_t\end{bmatrix}_t^T\right)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} Q & S \\ S^T & R\end{bmatrix}\right)$ *where* Q *is a* $p \times p$ *covariance matrix.*

(ii) $(\mathbf{u}_t)_{t\in\mathbb{N}}$ *is an* r*-dimensional exogenous input series and* A_t *a* $p \times r$ *matrix of parameters, which is possibly the zero matrix.*

(iii) The initial state $\mathbf{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_0)$*.*

(iv) Ψ_t *is a* $q \times p$ measurement *or* observation matrix *for all* $t \geq 1$*,*

(v) The matrix B_t *is a* $q \times r$ *regression matrix which may be the zero matrix.*

(vi) The initial state $\mathbf{X}_0$ *and the noise sequence* $((\mathbf{W}_t, \mathbf{V}_t)_t)_{t\in\mathbb{N}}$ *are independent.*

Following these changes in the model assumptions, Algorithm 6 has to be adapted as follows.

Algorithm 10: Kalman filter algorithm for correlated errors.

Data: Parameters Q, Θ_t, S, R and A_t, B_t, Ψ_t for $t = 1, \ldots, n$, initial conditions $\boldsymbol{\mu}$ and Σ_0, observations $\mathbf{Y}_t$ and exogenous input series $\mathbf{u}_t$, for $t = 1, \ldots, n$.

Result: Forecasting and filtering outputs $\mathbf{X}_{t|t-1}$, $\mathbf{X}_{t|t}$, and their autocovariance matrices $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$ for $t = 1, \ldots, n$.

Initialization: set $\mathbf{X}_{1|0} = \Phi_0\boldsymbol{\mu}\Phi_0^T + \mathrm{A}_1\mathbf{u}_1$ and $\boldsymbol{\Sigma}_{1|0} = \Phi_0\Sigma_0\Phi_0^T + \Theta_0 Q \Theta_0^T$.

for $t = 1, 2, \ldots, n$ **do**

Compute in this order

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t \tag{7.28}$$

$$\Gamma_t = \Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R, \tag{7.29}$$

$$K_t = [\Phi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + \Theta_t S]\Gamma_t^{-1}, \tag{7.30}$$

$$\mathbf{X}_{t+1|t} = \Phi_t \mathbf{X}_{t|t-1} + \mathrm{A}_{t+1}\mathbf{u}_{t+1} + K_t \boldsymbol{\epsilon}_t, \tag{7.31}$$

$$\boldsymbol{\Sigma}_{t+1|t} = \Phi_t \boldsymbol{\Sigma}_{t|t-1} \Phi_t^T + \Theta_t Q \Theta_t^T - K_t \Gamma_t^{-1} K_t^T, \tag{7.32}$$

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T \Gamma_t^{-1} \boldsymbol{\epsilon}_t, \tag{7.33}$$

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1} \Psi_{t+1}^T \Gamma_t^{-1} \Psi_t \boldsymbol{\Sigma}_{t|t-1}. \tag{7.34}$$

end

In this algorithm, $\boldsymbol{\epsilon}_t$ and Γ_t still correspond to the innovation process and its covariance matrix,

$$\boldsymbol{\epsilon}_t \overset{\text{def}}{=} \mathbf{Y}_t - \mathbf{Y}_{t|t-1}$$

$$\Gamma_t \overset{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T\right] = \mathrm{Cov}(\boldsymbol{\epsilon}_t) \,.$$

The adaptation of the proof of Proposition 7.2.1 to the correlated errors case is left to the reader (Exercise 7.3). The following result follows.

Proposition 7.4.1 (Kalman Filter for correlated errors)**.** *Algorithm 10 applies for the state-space model satisfying Assumption 7.4.1, provided that* $\Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R$ *are invertible matrices for* $t = 1, \ldots, n$.

7.5 Vector ARMAX models

Vector ARMAX models are a generalization of ARMA models to the case where the process is vector-valued and an eXternal input series is added to the model equation. Namely $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$ satisfies the following equation

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \sum_{j=1}^{p} \Phi_j \mathbf{Y}_{t-j} + \sum_{k=1}^{q} \Theta_k \mathbf{V}_{t-k} + \mathbf{V}_t \ . \tag{7.35}$$

The observations $\mathbf{Y}_t$ are a k-dimensional vector process, the Φs and Θs are $k \times k$ matrices, A is $k \times r$, $\mathbf{u}_t$ is the $r \times 1$ input, and $\mathbf{V}_t$ is a $k \times 1$ white noise process. The following result shows that such a model satisfies Assumption 7.4.1, under the additional Gaussian assumption. The proof is left to the reader (Exercise 7.4).

Proposition 7.5.1 (A State-Space Form of ARMAX)**.** *For* $p \geq q$*, let*

$$\Phi = \begin{bmatrix} \Phi_1 & \mathbf{1} & 0 & \cdots & 0 \\ \Phi_2 & 0 & \mathbf{1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_{p-1} & 0 & 0 & \cdots & \mathbf{1} \\ \Phi_p & 0 & 0 & \cdots & 0 \end{bmatrix} \quad \Theta = \begin{bmatrix} \Theta_1 + \Phi_1 \\ \vdots \\ \Theta_q + \Phi_q \\ \Phi_{q+1} \\ \vdots \\ \Phi_p \end{bmatrix} \quad \mathrm{A} = \begin{bmatrix} \mathrm{B} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where Φ *is* $kp \times kp$*,* Θ *is* $kp \times k$*,* B *is* $kp \times r$ *and* $\mathbf{1}$ *is the identity matrix (with adapted dimension depending on the context). Then, the state-space model given by*

$$\mathbf{X}_{t+1} = \Phi \mathbf{X}_t + \mathrm{A}\mathbf{u}_{t+1} + \Theta \mathbf{V}_t, \tag{7.36}$$

$$\mathbf{Y}_t = \Psi \mathbf{X}_t + \mathbf{V}_t, \tag{7.37}$$

where $\Psi = \begin{bmatrix} \mathbf{1}, 0, \cdots, 0 \end{bmatrix}$ *is* $k \times pk$*, implies the ARMAX model (7.35). If* $p < q$*, set* $\Phi_{p+1} = \cdots = \Phi_q = 0$*, and replace the value of* p *by that of* q *and (7.36)–(7.37) still apply. Note that the state process is* kp*-dimensional, whereas the observations are* k*-dimensional.*

Example 7.5.1 (ARMA(1, 1) with linear trend)**.** *Consider the univariate* $ARMA(1,1)$ *model with an additive linear trend*

$$Y_t = \beta_0 + \beta_1 t + \phi Y_{t-1} + \theta V_{t-1} + V_t \ .$$

Using Proposition 7.5.1, we can write the model as

$$X_{t+1} = \phi X_t + \beta_0 + \beta_1 t + (\theta + \phi) V_t, \tag{7.38}$$

and

$$Y_t = X_t + V_t. \tag{7.39}$$

Remark 7.5.1. *Since ARMA models are a particular case of DLM, the maximum likelihood estimation for Gaussian ARMA models can be performed using this general framework.*

Example 7.5.2 (Regression with autocorrelated errors)**.** *The (multivariate) regression with autocorrelated errors, is the regression model*

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t \ , \tag{7.40}$$

where we observe the $k \times 1$ vector-valued time series $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$ and $r \times 1$ regression-vectors $\mathbf{u}_t$, and where $(\boldsymbol{\epsilon}_t)_{t\in\mathbb{Z}}$ is a vector ARMA(p, q) process and B *is an unknown $k \times r$ matrix of regression parameters.*

This model is not an ARMAX because the regression is separated from the ARMA recursion. However, by proceeding as previously, it can also be defined in a state-space form. Using $\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \mathrm{B}\mathbf{u}_t$ is a k-dimensional ARMA(p, q) process, we have

$$\mathbf{X}_{t+1} = \Phi\mathbf{X}_t + \Theta\mathbf{V}_t, \tag{7.41}$$

$$\mathbf{Y}_t = \Psi\mathbf{X}_t + \mathrm{B}\mathbf{u}_t + \mathbf{V}_t, \tag{7.42}$$

where the model matrices Φ, Θ, and Ψ are defined in Proposition 7.5.1.

7.6 Likelihood of dynamic linear models

The dynamic linear model of Assumption 7.1.1 rely on a lot of parameters, namely $(\Psi_t)_{t\geq 1}$, $(\mathrm{A}_t)_{t\geq 1}$, Q, $\boldsymbol{\mu}_0$, Σ_0, $(\Phi_t)_{t\geq 1}$, $(\mathrm{B}_t)_{t\geq 1}$, and R, among which some or all entries may be unknown. We now consider the problem of estimating the unknown parameters of the dynamic linear model. Throughout this section, we suppose that Assumption 7.1.1 holds and moreover that the unknown parameters are not evolving with time. We denote by θ^* a vector containing all the unknown entries, or more generally speaking a given parameterization of the above original parameters. That is, to sum up, the framework in this section is the following.

1- The "original parameters" will be written as $(\Psi_t(\theta))_{t\geq 1}$, $(\mathrm{A}_t(\theta))_{t\geq 1}$, $Q(\theta)$, $\boldsymbol{\mu}_0(\theta)$, $\Sigma_0(\theta)$, $(\Phi_t(\theta))_{t\geq 1}$, $(\mathrm{B}_t(\theta))_{t\geq 1}$, and $R(\theta)$ with θ running through a given finite dimensional parameter set Θ and with θ^* denoting the true parameter used to generate the data (assuming that such a parmeter exists!).

2- As a result, each $\theta \in \Theta$ defines a precise (Gaussian) distribution for the observed data $\mathbf{Y}_{1:n}$.

3- It should be stressed that, although they could be quite helpful for estimating θ^*, the variables $\mathbf{X}_{1:n}$ are *unobserved*: one says that they are *hidden variables.*

In the following, we adapt the notation introduced in 7.2 to the Item 2- above. Namely, all quantities depending on the joint distribution of $\mathbf{X}_t, \mathbf{Y}_t$, $t = 1, \ldots, n$ can now be defined as function of $\theta \in \Theta$. For instance, Equations (7.12) and (7.13) become

$$\boldsymbol{\epsilon}_t(\theta) = \mathbf{Y}_t - \Psi_t(\theta)\mathbf{X}_{t|t-1}(\theta) - \mathrm{B}_t(\theta)\mathbf{u}_t, \tag{7.43}$$

$$\Gamma_t(\theta) = \Psi(\theta)\boldsymbol{\Sigma}_{t|t-1}(\theta)\Psi_t(\theta)^T + R(\theta) \tag{7.44}$$

Here $\mathbf{X}_{t|t-1}(\theta)$ and $\boldsymbol{\Sigma}_{t|t-1}(\theta)$ also depend on θ since they are now functions of the parameter θ which determines the joint distribution of the hidden and observed data $\mathbf{X}_t, \mathbf{Y}_t$, $t = 1, \ldots, n$.

Based on the general equations (4.35), (4.38) and (4.39), in the framework of a parameterized dynamic linear model, (4.40) thus gives that

$$-2\log L_n(\theta) = n\log(2\pi) + \sum_{t=1}^{n} \log\det\Gamma_t(\theta) + \sum_{t=1}^{n} \boldsymbol{\epsilon}_t(\theta)^T \Gamma_t(\theta)^{-1} \boldsymbol{\epsilon}_t(\theta)\ , \tag{7.45}$$

provided that $\Gamma_t(\theta)$ is invertible for all $t = 1, \ldots, n$ and $\theta \in \Theta$.

Observe that, for each θ, the negated log likelihood $-2\log L_n(\theta)$ can thus be efficiently computed by running the Kalman filter (see Algorithm 6) and then applying (7.43), (7.44) and (7.45).

Similarly one can compute the gradient $-\partial \log L_n(\theta)$ and the Hessian $-\partial\partial^T \log L_n(\theta)$, provided that the original parameters are at least twice differentiable with respect to θ. Formula (4.41) and (4.42) can directly be applied replacing $\boldsymbol{\eta}$ by $\boldsymbol{\epsilon}$ and $\widetilde{\Sigma}$ by Γ.

However one needs to adapt Algorithm 6 to compute the gradient or the Hessian. A rather simple case is obtained when the Ψ_ts are known design matrices (that is, they do not depend on θ). In this case differentiating within Algorithm 6 provides the following algorithm.

Algorithm 11: Kalman filter algorithm for the gradient of the likelihood.

Data: A parameter $\theta \in \Theta$, observations $\mathbf{Y}_t$ and exogenous input series $\mathbf{u}_t$, for $t = 1, \ldots, n$, an index i. The functions and their first derivatives Q, R, A_t, B_t, Ψ_t for $t = 1, \ldots, n$, $\boldsymbol{\mu}$ and Σ_0 can be evaluated at θ. Functions K_t, $\mathbf{X}_{t|t-1}$, $\mathbf{X}_{t|t}$, $\boldsymbol{\Sigma}_{t|t-1}$, $\boldsymbol{\Sigma}_{t|t}$, Γ_t and $\boldsymbol{\epsilon}_t$ are already computed at θ for $t = 1, \ldots, n$.

Result: i-th component of the forecasting errors' gradient $\partial_i \boldsymbol{\epsilon}_t(\theta)$ and error covariance gradient $\partial_i \Gamma_t(\theta)$ at θ.

Initialization: set $\partial_i \mathbf{X}_{0|0}(\theta) = \partial_i \boldsymbol{\mu}_0(\theta)$ and $\partial_i \boldsymbol{\Sigma}_{0|0}(\theta) = \partial_i \Sigma_0(\theta)$.

for $t = 1, 2, \ldots, n$ **do**

Compute in this order (the following functions are evaluated at θ)

$$\begin{aligned}
\partial_i \mathbf{X}_{t|t-1} &= [\partial_i \Phi_t]\mathbf{X}_{t-1|t-1} + \Phi_t[\partial_i \mathbf{X}_{t-1|t-1}] + [\partial_i \mathrm{A}_t]\mathbf{u}_t, \\
\partial_i \boldsymbol{\Sigma}_{t|t-1} &= [\partial_i \Phi_t]\boldsymbol{\Sigma}_{t-1|t-1}\Phi_t^T + \Phi_t[\partial_i \boldsymbol{\Sigma}_{t-1|t-1}]\Phi_t^T \\
&\qquad + \Phi_t \boldsymbol{\Sigma}_{t-1|t-1}[\partial_i \Phi_t]^T + \partial_i Q, \\
\partial_i \boldsymbol{\epsilon}_t &= -\Psi_t[\partial_i \mathbf{X}_{t|t-1}] - [\partial_i \mathrm{B}_t]\mathbf{u}_t, \\
\partial_i \Gamma_t &= \Psi_t[\partial_i \boldsymbol{\Sigma}_{t|t-1}]\Psi_t^T + \partial_i R(\theta) \\
\partial_i K_t &= \left\{[\partial_i \boldsymbol{\Sigma}_{t|t-1}]\Psi_t^T - K_t[\partial_i \Gamma_t]\right\}\Gamma_t^{-1}. \\
\partial_i \mathbf{X}_{t|t} &= [\partial_i \mathbf{X}_{t|t-1}] + [\partial_i K_t]\boldsymbol{\epsilon}_t + K_t[\partial_i \boldsymbol{\epsilon}_t], \\
\boldsymbol{\Sigma}_{t|t} &= [\partial_i K_t]\Psi_t \boldsymbol{\Sigma}_{t|t-1} + [I - K_t\Psi_t][\partial_i \boldsymbol{\Sigma}_{t|t-1}].
\end{aligned}$$

end

Algorithm 6 and Algorithm 11 can be used with a gradient descent type numerical algorithm that provides a numerical approximation of the minimizer of $\theta \mapsto -\log L_n(\theta)$.

(i) Select initial values for the parameters, say, $\theta^{(0)}$.

(ii) Run the Kalman filter, Proposition 7.2.1, using the initial parameter values, $\theta^{(0)}$, to obtain a set of innovations and error covariances, say, $\{\boldsymbol{\epsilon}_t^{(0)};\ t = 1, \ldots, n\}$ and $\{\Gamma_t^{(0)};\ t = 1, \ldots, n\}$.

(iii) Run one iteration of a Newton–Raphson procedure with $-\log L_Y(\theta)$ as the criterion function to obtain a new set of estimates, say $\theta^{(1)}$.

(iv) At iteration j, $(j = 1, 2, \ldots)$, repeat step 2 using $\theta^{(j)}$ in place of $\theta^{(j-1)}$ to obtain a new set of innovation values $\{\boldsymbol{\epsilon}_t^{(j)};\ t = 1, \ldots, n\}$ and $\{\Gamma_t^{(j)};\ t = 1, \ldots, n\}$. Then repeat step 3 to obtain a new estimate $\theta^{(j+1)}$. Stop when the estimates or the likelihood stabilize.

Example 7.6.1 (Noisy AR(1) (continued from Example 7.1.3 and Example 7.2.1)**.** *Let us apply a standard numerical procedure*[1] *to compute estimates of the parameter* $\theta = (\phi, \sigma_w^2, \sigma_v^2)$ *from a simulated samples of Example 7.1.3 with length* $n = 128$*. We replicate this experiment for fixed parameters* $\phi = 0.8$ *and* $\sigma_v = 1.0$ *and* $\sigma_w = 1.0$*. The distribution of the obtained estimates are displayed using boxplots in Figure 7.5.*

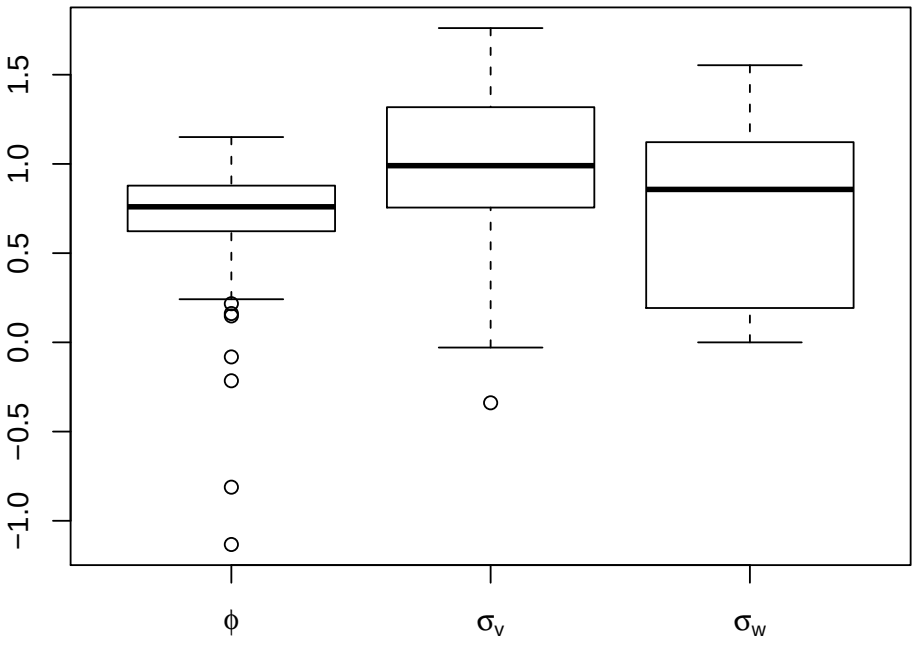

Figure 7.5: Estimation of the parameters of the noisy AR(1) model: boxplots of the estimates of ϕ, σ_v and σ_w obtained from 100 Monte Carlo replications of time series of length 128. The true values are $\phi = 0.8$ and $\sigma_v = 1.0$ and $\sigma_w = 1.0$.

To conclude, we mention that the EM algorithm introduced in Section 4.8 is very well adapted to the dynamic linear model, where the hidden variables $\mathbf{U}_n$ are taken as $\mathbf{U}_n = \mathbf{X}_{0:n}$. Let us derive the Expectation and the minimization steps in the particular (and a little bit

[1] the quasi Newton procedure implemented in the `optim()` function of the R software, R software

simpler) case where the input sequence $(\mathbf{u}_t)$ is absent. In this case, using Assumption 7.1.1, the joint likelihood reads

$$p(\mathbf{X}_{0:n}, \mathbf{Y}_{1:n}|\theta) = p(\mathbf{X}_0|\theta) \prod_{t=1}^{n} p(\mathbf{X}_t|\mathbf{X}_{t-1}, \theta) \prod_{t=1}^{n} p(\mathbf{Y}_t|\mathbf{X}_t, \theta) \ .$$

Moreover the above densities are all Gaussian and thus, up to additive constants, $-2\log p(\mathbf{X}_0|\theta)$, $-2\log p(\mathbf{X}_t|\mathbf{X}_{t-1}, \theta)$ and $-2\log p(\mathbf{Y}_t|\mathbf{X}_t, \theta)$ are respectively equal to

$$\begin{aligned}
&\log\det\Sigma_0(\theta) + (\mathbf{X}_0 - \boldsymbol{\mu}_0(\theta))^T \Sigma_0^{-1}(\theta)(\mathbf{X}_0 - \boldsymbol{\mu}_0(\theta)) \ , \\
&\log\det Q(\theta) + (\mathbf{X}_t - \Phi_t(\theta)\mathbf{X}_{t-1})^T Q^{-1}(\theta)(\mathbf{X}_t - \Phi_t(\theta)\mathbf{X}_{t-1}) \ , \\
&\log\det R(\theta) + (\mathbf{Y}_t - \Psi_t(\theta)\mathbf{X}_t)^T R^{-1}(\theta)(\mathbf{Y}_t - \Psi_t(\theta)\mathbf{X}_t) \ .
\end{aligned}$$

Computing the conditional expectation given the observed data $\mathbf{Y}_{1:n}$ yields

$$\begin{aligned}
\mathcal{Q}_n(;\theta\,;\theta') = &\log\det\Sigma_0 + \mathrm{Trace}\left\{\Sigma_0^{-1}[\boldsymbol{\Sigma}_{0|n} + (\mathbf{X}_{0|n} - \boldsymbol{\mu})(\mathbf{X}_{0|n} - \boldsymbol{\mu})^T]\right\} \\
&+ n\log\det Q + \mathrm{Trace}\left\{Q^{-1}[S_{00} - S_{01}\Phi^T - \Phi S_{10} + \Phi S_{11}\Phi^T]\right\} \\
&+ n\log\det R + \mathrm{Trace}\Big\{R^{-1}\sum_{t=1}^{n}[(\mathbf{Y}_t - \Psi_t\mathbf{X}_{t|n})(\mathbf{Y}_t - \Psi_t\mathbf{X}_{t|n})^T + \Psi_t\boldsymbol{\Sigma}_{t|n}\Psi_t^T]\Big\},
\end{aligned}$$

where, for $k, \ell = 0, 1$,

$$S_{k,\ell} = \sum_{t=1}^{n}(\mathbf{X}_{t-k|n}\mathbf{X}_{t-\ell|n}^T + \boldsymbol{\Sigma}_{t-k,t-\ell|n}) \ .$$

In these equations, the quantity relative to the smoothing are calculated under the parameter θ', while the other parameters are functions of θ. We have not explicitly displayed this fact, for the sake of clarity.

Let us consider the case where Ψ_t is known for all t and $\Phi_t = \Phi$, so that the unknowns are $\theta = (\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q, R)$. We can then minimize $\mathcal{Q}_n(;\theta\,;\theta')$ above with respect to θ as in the usual multivariate regression approach, which yields the updated parameters

$$\begin{aligned}
\boldsymbol{\mu}_0 &= \mathbf{X}_{0|n} \\
\Sigma_0 &= \boldsymbol{\Sigma}_{0|n} \\
\Phi &= S_{01}S_{11}^{-1}, \\
Q &= n^{-1}\left(S_{00} - S_{01}S_{11}^{-1}S_{10}\right), \\
R &= n^{-1}\sum_{t=1}^{n}[(\mathbf{Y}_t - \Psi_t\mathbf{X}_{t|n})(\mathbf{Y}_t - \Psi_t\mathbf{X}_{t|n})^T + \Psi_t\boldsymbol{\Sigma}_{t|n}\Psi_t^T] \ .
\end{aligned}$$

Example 7.6.2 (Noisy AR(1) (continued from Example 7.1.3, Example 7.2.1 and Example 7.6.1))**.** *Let us use the EM algorithm to obtain a numerical sequence approaching the MLE whose corresponding likelihood sequence increases at each step. The estimation result is similar to that using a standard numerical procedure as in Example 7.6.1. In Figure 7.6, the log likelihood at each iteration is displayed.*

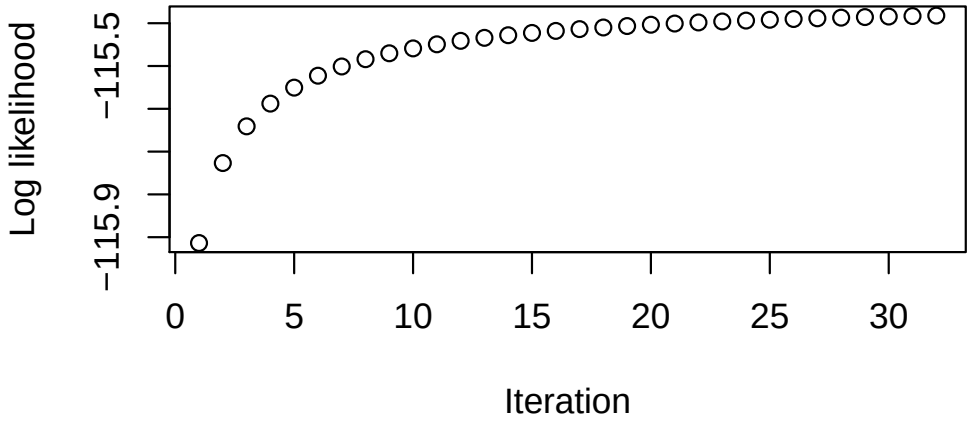

Figure 7.6: Estimation for the parameters of the noisy AR(1) model using the EM algorithm: log likelihood at each iteration.

7.7 Exercises

Exercise 7.1. Show that the process $(Y_t)_{t\in\mathbb{Z}}$ of Example 7.1.3 is an ARMA(1,1) process.

Exercise 7.2. Show that Algorithm 9 applies under the assumptions of Proposition 7.2.1.

Exercise 7.3. Show that Algorithm 10 applies for the state-space model satisfying Assumption 7.4.1, provided that $\Psi_t \mathbf{\Sigma}_{t|t-1} \Psi_t^T + R$ are invertible matrices for $t = 1, \dots, n$.

Exercise 7.4. Prove Proposition 7.5.1.

Exercise 7.5 (Kalman filtering of an AR(1) process observed in noise). Consider an AR(1) process (X_t) with canonical representation:

$$X_{t+1} = \phi X_t + W_{t+1} \tag{7.46}$$

where (W_t) is a centered white noise with known variance σ^2 and ϕ is also known. The process (X_t) is not directly observable and we have, for $t \geq 1$,

$$Y_t = X_t + V_t\,, \tag{7.47}$$

where (V_t) is a centered white noise with known variance ρ^2, that is uncorrelated with (W_t).

We denote by $\hat{X}_{t|t} = \text{proj}\left(X_t \,\middle|\, \mathcal{H}_{t,t}^Y\right)$ the filtering estimate and by $\Sigma_{t|t} = \mathbb{E}(X_t - \hat{X}_{t|t})^2$ the corresponding projection error variance. Similarly, $\hat{X}_{t+1|t} = \text{proj}\left(X_{t+1} \,\middle|\, \mathcal{H}_{t,t}^Y\right)$ is the best linear state predictor and $\Sigma_{t+1|t} = \mathbb{E}(X_{t+1} - \hat{X}_{t+1|t})^2$ the corresponding error variance.

1. Comment the differences between the model of Equations (7.46)–(7.47) and the general state-space representation.

2. Using the evolution equation (7.46), show that

$$\hat{X}_{t+1|t} = \phi \hat{X}_{t|t} \quad \text{and} \quad \Sigma_{t+1|t} = \phi^2 \Sigma_{t|t} + \sigma^2$$

3. Defining the innovation[2] by $I_{t+1} = Y_{t+1} - \text{proj}\left(Y_{t+1} \,\middle|\, \mathcal{H}^Y_{t,t}\right)$. Using the observation equation (7.47), show that $I_{t+1} = Y_{t+1} - \hat{X}_{t+1|t}$.

4. Prove that $\mathbb{E}[I^2_{t+1}] = \Sigma_{t+1|t} + \rho^2$.

5. Why does the following expression holds:

$$\hat{X}_{t+1|t+1} = \hat{X}_{t+1|t} + k_{t+1} I_{t+1}\,,$$

where $k_{t+1} = \mathbb{E}[X_{t+1} I_{t+1}]/\mathbb{E}[I^2_{t+1}]$ (k_{t+1} is called *the Kalman gain*)?

6. Using the above expression of I_{t+1}, show that $\mathbb{E}[X_{t+1} I_{t+1}] = \Sigma_{t+1|t}$.

7. Show that $\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - \mathbb{E}\left[(k_{t+1} I_{t+1})^2\right]$ and deduce from this that $\Sigma_{t+1|t+1} = (1 - k_{t+1})\Sigma_{t+1|t}$.

8. Provide the complete set of equations that allows to compute $\hat{X}_{t|t}$ and $\Sigma_{t|t}$ iteratively for all $t \geq 1$.

9. Study the asymptotic behavior of $\Sigma_{t|t}$ as $t \to \infty$.

[2] Using the notation of Section 2.7.2, this corresponds to $\epsilon^+_{t+1,t}$.

Chapter 8

Non-stationary time series

8.1 Limitations of weakly stationary ARMA modelling

We have seen in Chapters 5 and 7 models that extend the class of ARMA models in several directions. These models are defined through iterative equations. In Chapter 7, we even restrict ourselves to deterministic linear equations why the linearity in stochastic autoregressive models of Chapter 5 is driven by possibly random matrices. Also these extensions included the possibility of vector valued processes. We will come back to the vector valued case again in Chapter 9, this time by focusing on particular non-stationary behaviors specific to multivariate time series with a vector AR (VAR) structure. It is also important to note that the iterative models introduced in Chapter 5 are only the tip of the iceberg of a much more general class of models, the *partially observed Markov chains*, which include

a) the *hidden Markov models*, whose linear subclass exactly corresponds to the dynamic linear models of Chapter 7 but which also include the stochastic volatility model of Chapter 5;

b) or the *observation driven models*, which contain the very popular GARCH models of Chapter 5 but also many extensions of the GARCH models including the discrete time series models (or time series of counts) such as GARCH Poisson processes.

All these models are generally studied in a stationary context and share a very nice property with the ARMA processes: their time correlations (when they have one) typically decrease exponentially fast; in the Markov theory context, one says that the Markov chain is *geometrically ergodic*. Of course it is possible to build non linear Markov chains that do not enjoy this property but, they are rarely used in practice for time series modeling. The reason is not that practitioners (including those dealing with financial time series) are not interested in slowly decreasing correlations but they generally do not wish to add difficulties on top of each other; hence, in the context of slowly decreasing correlations, linear time series are mostly considered, that is, time series with representations of the form

$$X_t = \sum_{k\in\mathbb{Z}} \psi_k \epsilon_{t-k} \ , \tag{8.1}$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ is a strong white noise, sometimes only starting at 0, that is, we will set $\epsilon_t = 0$ for $t \leq 0$. ARMA processes typically have this representation (see Theorem 3.3.2), however

with ψ_k exponentially decreasing as $|k| \to \infty$, which indeed implies the exponential decreasing of the time correlation. Let us recall an important example (already introduced, see Example 2.3.3) which does not fit this behavior.

Example 8.1.1 (Random walk). *Let $(\epsilon_t)_{t\in\mathbb{N}}$ be a strong white noise and define, for all $t \in \mathbb{N}$,*

$$X_t = \sum_{0<k\leq t} \epsilon_k \ ,$$

which corresponds to (8.1) with $\psi_k = 1$ for all $k \in \mathbb{Z}$ and setting $\epsilon_t = 0$ for $t \leq 0$. Then we have, for all $t \in \mathbb{N}^*$,

$$\frac{\mathrm{Cov}\,(X_t, X_0)}{\sqrt{\mathrm{Var}\,(X_t)\,\mathrm{Var}\,(X_0)}} = t^{-1/2} \ .$$

In fact, the random walk example is somewhat an extreme case, since it is not even a weakly stationary process. It can be seen as a non-stationary AR(1) process since it satisfies the equation

$$X_{t+1} = X_t + \epsilon_{t+1} \ , \qquad t \in \mathbb{N} \ .$$

Hence, in some sense it can be compared to an AR(1) stationary process with AR coefficient ϕ close to 1. It is based on this remark that the unit root test will be constructed in Section 8.7.

Between the nonstationary random walk and the stationary ARMA process, one can exhibit a class of random processes which are weakly stationary but whose correlation has a slower decay than the exponential one of ARMA processes, as will be done Section 8.3, the obtained processes are said to exhibit *long range dependence* (LRD) or *long memory*.

To introduce LRD processes, we will need to rely on particular filters that can be suitably introduced using spectral representations, as shown in Section 8.2. Finally, we will conclude this chapter with the parametric class of FARIMA processes, which the econometricians Clive Granger and Roselyne Joyeux have introduced in Granger and Joyeux [1980].

8.2 Linear filtering via spectral representation

8.2.1 Definition

Let us start with a simple example where, given a weakly stationary process X and a random variable in $\mathcal{H}_\infty^X$, one obtain a linear filter.

Example 8.2.1 (Linear filtering in $\mathcal{H}_\infty^X$). *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a centered a weakly stationary process with autocovariance γ and let $Y_0 \in \mathcal{H}_\infty^X$. Then there exists an array of complex numbers $(\alpha_{n,s})_{s\in\mathbb{Z},n\geq 1}$ such that for all $n \in \mathbb{N}$, the set $\{s \in \mathbb{Z}, \alpha_{n,s} \neq 0\}$ is finite and, as $n \to \infty$,*

$$\sum_{s\in\mathbb{Z}} \alpha_{n,s} X_{-s} \to Y_0 \quad \textit{in } L^2.$$

It follows that, by weak stationarity and using the Cauchy criterion, for all $t \in \mathbb{Z}$,

$$\sum_{s\in\mathbb{Z}} \alpha_{n,s} X_{t-s} \to Y_t \quad \textit{in } L^2,$$

where $Y_t \in \mathcal{H}_\infty^X$. By continuity of the expectation and the scalar product, we easily obtain that the process $Y = (Y_t)_{t\in\mathbb{Z}}$ is a centered weakly stationary process with autocovariance function

$$\gamma'(\tau) = \lim_{n\to\infty} \sum_{s\in\mathbb{Z}} \sum_{t\in\mathbb{Z}} \alpha_{n,s}\alpha_{n,t}\gamma(\tau - t + s) \ .$$

A particular case of the previous case is obtained when X is a white noise.

Example 8.2.2 (The white noise case)**.** *We consider Example 8.2.1 with* $X \sim \mathrm{WN}(0, \sigma^2)$. *In this case* $(X_t)_{t\in\mathbb{Z}}$ *is a Hilbert basis of* $\mathcal{H}_\infty^X$ *and thus*

$$\mathcal{H}_\infty^X = \left\{ \sum_{t\in\mathbb{Z}} \alpha_t X_t \,:\, (\alpha_t) \in \ell^2(\mathbb{Z}) \right\} ,$$

where $\ell^2(\mathbb{Z})$ *is the set of sequences* $(x_t) \in \mathbb{C}^{\mathbb{Z}}$ *such that* $\sum_t |\alpha_t|^2 < \infty$ *and the convergence of* $\sum_{t\in\mathbb{Z}}$ *is understood in the* L^2 *sense. As a result we may take* $(\alpha_{n,t})_{t\in\mathbb{Z},n\geq 1}$ *as* $\alpha_{n,t} = \alpha_t \mathbb{1}(-n \leq t \leq n)$ *and obtain*

$$Y_t = \sum_{s\in\mathbb{Z}} \alpha_s X_{t-s} \quad \text{in } L^2, \tag{8.2}$$

and

$$\gamma'(\tau) = \sum_{s\in\mathbb{Z}} \sum_{t\in\mathbb{Z}} \alpha_s \alpha_t \gamma(\tau - t + s) \, .$$

Unfortunately, it is not always possible to choose $(\alpha_{n,t})_{t\in\mathbb{Z},n\geq 1}$ as in Example 8.2.2, that is, of the form $\alpha_{n,t} = \alpha_t \mathbb{1}(-n \leq t \leq n)$ for some sequence $(\alpha_t)_{t\in\mathbb{Z}}$. When (8.2) does hold for all t, we will use the notation introduced in (3.1) and write

$$Y = \mathrm{F}_\alpha(X) \, .$$

When it dos not, using a converging array $(\alpha_{n,t})_{t\in\mathbb{Z},n\geq 1}$ to define Y_t may appear complicated. Nevertheless, the spectral representation of Y_0 provides a very helpful simplification.

Let us consider again the general approach of Example 8.2.1. Then Y_0 is uniquely determined by its spectral representation, see Theorem 2.5.6,

$$Y_0 = \int g(\lambda) \, \mathrm{d}\hat{X}(\lambda) \, ,$$

where $\hat{X}$ is the spectral field of X with intensity measure ν, the spectral measure of X, and $g \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$. In particular, by the unitary property, we have

$$\sum_{s\in\mathbb{Z}} \alpha_{n,s} \mathrm{e}^{-\mathrm{i}\lambda s} \to g(\lambda) \quad \text{in } L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu),$$

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

$$Y_t = \int \mathrm{e}^{\mathrm{i}\lambda t} \, g(\lambda) \, \mathrm{d}\hat{X}(\lambda) \, ,$$

With this formulation of Example 8.2.1, we directly obtain that Y has spectral field $\mathrm{d}\hat{Y}(\lambda) = g(\lambda) \, \mathrm{d}\hat{X}(\lambda)$ and spectral measure $\mathrm{d}\nu'(\lambda) = |g(\lambda)|^2 \, \mathrm{d}\nu(\lambda)$. This point of view can formally be expressed by the following statements.

Definition 8.2.1 (Density of a random field with orthogonal increments)**.** *Let* V *and* W *be random fields with orthogonal increments on* $(\mathbb{X}, \mathcal{X})$ *with intensity measure* η *and* ξ. *Let moreover* $\alpha \in L^2(\mathbb{X}, \mathcal{X}, \xi)$. *We say that* V *has density* α *with respect to* W *if, for all* $f \in L^2(\mathbb{X}, \mathcal{X}, \eta)$,

$$\int f \mathrm{d}V = \int f \, \alpha \, \mathrm{d}W \, .$$

We will write $\mathrm{d}V = \alpha \, \mathrm{d}W$. *Then* η *has density* $|\alpha|^2$ *with respect to* ξ.

Proposition 8.2.1. *Let W be a random field with orthogonal increments on $(\mathbb{X}, \mathcal{X})$ with intensity measure ξ. Let moreover $\alpha \in L^2(\mathbb{X}, \mathcal{X}, \xi)$. Then there exists a unique random field V with orthogonal increments on $(\mathbb{X}, \mathcal{X})$ that has density α with respect to W.*

Proof. Clearly the operator $f \to \alpha \times f$ is a unitary operator from $L^2(\mathbb{X}, \mathcal{X}, \eta)$ to $L^2(\mathbb{X}, \mathcal{X}, \xi)$ where η has density $|\alpha|^2$ with respect to ξ. By Theorem 2.5.3, we deduce that $f \mapsto \int (f \times \alpha) \mathrm{d}W$ is a unitary operator from $L^2(\mathbb{X}, \mathcal{X}, \eta)$ to $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, using Theorem 2.5.4, we conclude the proof. □

The last statement sum up the above construction.

Theorem 8.2.2. *Let X be a centered a weakly stationary process with spectral field $\hat{X}$ and spectral measure ν. Let $\alpha \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$. Then the process $Y = (Y_t)_{t \in \mathbb{Z}}$ defined by*

$$Y_t = \int \mathrm{e}^{\mathrm{i}t\lambda} \alpha(\lambda) \, \mathrm{d}\hat{X}(\lambda) \, , \quad t \in \mathbb{Z} \, , \tag{8.3}$$

is centered and weakly stationary. Moreover its spectral field $\hat{Y}$ has density α with respect to $\hat{X}$, $\mathrm{d}\hat{Y} = \alpha \mathrm{d}\hat{X}$ and its spectral measure has density $|\alpha|^2$ with respect to ν.

Definition 8.2.2 (Linear filtering in $\mathcal{H}_\infty^X$ with transfer function α)**.** *Under the assumptions of Theorem 8.2.2, we will say that Y is the output of the linear filter with transfer function α and input X. We will write*

$$Y = \widehat{\mathrm{F}}_\alpha(X) \, .$$

Of course this definition should be compared to the approach used in (3.1) for $\psi \in \ell^1$. In this case, for any centered weakly stationary process X we have

$$\mathrm{F}_\psi(X) = \widehat{\mathrm{F}}_\alpha(X)$$

where

$$\alpha(\lambda) = \sum_{s \in \mathbb{Z}} \psi_k \mathrm{e}^{-\mathrm{i}s\lambda} \, .$$

8.2.2 Composition and inversion via spectral representation

In Definition 8.2.2, the transfer function α must satisfy some condition depending on the spectral density ν of X. In turns, this implies that for a given function α on $\mathbb{T}$, the linear filtering $\widehat{\mathrm{F}}_\alpha(X)$ may not be correctly defined for any X. This question is crucial before considering the composition of two linear filtering.

We first need some notation.Let $\mathcal{S}(\Omega, \mathcal{F}, \mathbb{P})$ (or simply $\mathcal{S}$ if no ambiguity occurs) denote the set of all centered weakly stationary processes indexed on $\mathbb{Z}$, defined on $\mathcal{S}(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $\mathbb{C}$. For any finite measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$, we further denote by $\mathcal{S}_\nu(\Omega, \mathcal{F}, \mathbb{P})$ (or simply $\mathcal{S}_\nu$) the subset of $\mathcal{S}$ of all centered weakly stationary processes with spectral measure ν.

For a given measurable function α from $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ to $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$, we denote by $\mathcal{M}[\alpha]$ the set of all finite measures ν on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ such that

$$\int |\alpha|^2 \, \mathrm{d}\nu < \infty \, .$$

In particular if α is bounded, $\mathcal{M}[\alpha]$ contains all finite measures ν on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. It follows from the above notations that $\widehat{\mathrm{F}}_\alpha(X)$ is defined on the set

$$\{X \in \mathcal{S},\ \nu_X \in \mathcal{M}[\alpha]\}\ ,$$

where ν_X denotes the spectral measure of X. We thus get that for any given measurable functions α and β from $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ to $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$, $\widehat{\mathrm{F}}_\beta \circ \widehat{\mathrm{F}}_\alpha$ is defined on the set

$$\{X \in \mathcal{S},\ \nu_X \in \mathcal{M}[\alpha] \cap \mathcal{M}[\alpha\beta]\}\ .$$

This set may be different from the definition set of $\widehat{\mathrm{F}}_\alpha \circ \widehat{\mathrm{F}}_\beta$. This is obviously the case for $\beta = 1/\alpha$ with α bounded and β unbounded. A classical example is obtained with $\alpha(\lambda) = 1 - \mathrm{e}^{-\mathrm{i}\lambda}$ for $\lambda \neq 0$ and, say, $\alpha(0) = 1$. On the other hand, if one takes good care of this problem (although it is rarely the case in the literature), we get the following result.

Proposition 8.2.3. *Let α and β be two measurable functions from $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ to $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$. Let X be a centered weakly stationary process with spectral measure ν. Then the following assertions hold.*

(i) If $\nu \in \mathcal{M}[\alpha] \cap \mathcal{M}[\alpha\beta]$ we have

$$\widehat{\mathrm{F}}_\beta \circ \widehat{\mathrm{F}}_\alpha(X) = \widehat{\mathrm{F}}_{\alpha\times\beta}(X)\ .$$

(ii) If $\nu \in \mathcal{M}[\alpha] \cap \mathcal{M}[\alpha\beta] \cap \mathcal{M}[\beta]$ we have

$$\widehat{\mathrm{F}}_\beta \circ \widehat{\mathrm{F}}_\alpha(X) = \widehat{\mathrm{F}}_\alpha \circ \widehat{\mathrm{F}}_\beta(X) = \widehat{\mathrm{F}}_{\alpha\times\beta}(X)\ .$$

Proof. Clearly Assertion *(ii)* is a consequence of Assertion *(i)*.

Now, Assertion *(i)* can be obtained using the spectral representations of X and $Y = \widehat{\mathrm{F}}_\alpha(X)$ and their unitary properties. Details are left to the reader (see Exercise 8.1). □

We now provide a result for the inversion of a linear filter of the form $Y = \widehat{\mathrm{F}}_\alpha(X)$ as introduced in Definition 8.2.2.

Proposition 8.2.4. *Let X be a centered weakly stationary process with spectral field $\hat{X}$ and spectral measure ν and $\alpha \in L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$. Define the weakly stationary process $Y = \widehat{\mathrm{F}}_\alpha(X)$. Then $\mathcal{H}_\infty^Y \subset \mathcal{H}_\infty^X$. Moreover, if $\alpha > 0$ ν-a.e. then $\mathcal{H}_\infty^X = \mathcal{H}_\infty^Y$ and we have $X = \widehat{\mathrm{F}}_{1/\alpha}(Y)$.*

Proof. Let ν' be the spectral measure of Y ($\mathrm{d}\nu' = |\alpha|^2 \mathrm{d}\nu$). We proceed as in the proof of Proposition 8.2.1. Using that $A : f \mapsto \alpha \times f$ is a unitary operator from $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu')$ to $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$, by Theorem 2.5.6, we obtain a unitary operator from $\mathcal{H}_\infty^Y$ to $\mathcal{H}_\infty^X$. Hence $\mathcal{H}_\infty^Y \subset \mathcal{H}_\infty^X$.

Moreover this unitary operator is surjective if and only if A is surjective, that is, if $\alpha > 0$ ν-a.e., in which case A^{-1} is the operator $f \mapsto f/\alpha$. We thus get that $X = \widehat{\mathrm{F}}_{1/\alpha}(Y)$, which concludes the proof. □

Observe that Proposition 8.2.3 and Proposition 8.2.4 are generalizations of Lemma 3.2.1 and Proposition 3.2.2. The spectral approach can be used to solve the ARMA equations, obtaining similar results as in Section 3.3. In particular the ARMA equation of the form (3.25), or equivalently (3.26) can be written as

$$\widehat{\mathrm{F}}_{\phi^*}(X) = \widehat{\mathrm{F}}_{\theta^*}(Z)\ , \tag{8.4}$$

where ϕ^* and θ^* are defined by

$$\phi^*(\lambda) = \Phi(\mathrm{e}^{-\mathrm{i}\lambda}) \ ,$$

and

$$\theta^*(\lambda) = \Theta(\mathrm{e}^{-\mathrm{i}\lambda}) \ .$$

Applying Proposition 8.2.4, we immediately find that, if Φ does not vanish on the unit circle, then we can apply $\widehat{\mathrm{F}}_{1/\phi^*}$ to both sides of (8.4) to get that the unique solution of this equation is

$$X = \widehat{\mathrm{F}}_{1/\phi^*} \circ \widehat{\mathrm{F}}_{\theta^*}(Z) \ .$$

8.3 Fractional integration and long range dependence

There are many different definitions of long range dependence in the literature, which, unfortunately are not exactly equivalent. They all contain the same classical examples but one can exhibit examples that fit one definition but not another one. Here we will use the definition involving the spectral density, which is more common in the econometric literature since the works of Robinson [1995].

Definition 8.3.1 (Long range dependence (LRD))**.** *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a weakly stationary process admitting a spectral density f. We say that X is* long range dependent *with* long memory *parameter d if f has the form*

$$f(\lambda) = \left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{-2d} \ f^*(\lambda) \ , \tag{8.5}$$

where f^* is a spectral density which is continuous and positive at zero.

This definition is in fact equivalent to saying that $f(\lambda) \sim |\lambda|^{-2d} f^*(0)$ as $\lambda \to 0$, with $f^*(0) > 0$. If $d = 0$ we usually say that X is *short range dependent* and if $d < 0$ that it has *negative long memory*. Since the spectral density is integrable, we see that we must have

$$d < 1/2 \ .$$

In fact when extending the notion of LRD to nonstationary processes d is allowed to take any value in $\mathbb{R}$, see Section 8.5.

The reason behind the above peculiar form involving the function

$$\lambda \mapsto \left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{-2d} = |2(1 - \cos(\lambda))|^{-d} = 4^{-d} \ |\sin(\lambda/2)|^{-2d} \ .$$

is to interpret a process with spectral density (8.5) as the result of a particular filtering of a process with spectral density f^*. Let $Y = (Y_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process with spectral density f^*, assumed to be continuous and positive at zero, and spectral representation $\hat{Y}$. Define $\iota_d : \mathbb{T} \to \mathbb{C}$ by

$$\iota_d(\lambda) = \begin{cases} \left(1 - \mathrm{e}^{-\mathrm{i}\lambda}\right)^{-d} & \text{if } \lambda \neq 0 \text{ or } d \leq 0 \\ \mathrm{i}^{-d}\infty & \text{otherwise.} \end{cases} \tag{8.6}$$

Then

$$X = \widehat{\mathrm{F}}_{\iota_d}(Y)$$

is well defined if $d < 1/2$, and has spectral density (8.5). Observe that if $d = 0$ $X = Y$, and if d is a negative integer, then

$$X = (\mathbf{1} - \mathrm{B})^{-d} Y \ ,$$

is obtained by differencing Y $(-d)$-times, or in other words, X is obtained by integrating Y d times (with $d < 0$!). This terminology is extending to all $d \in \mathbb{R}$ in the following.

Definition 8.3.2 (Integrating, differencing operator)**.** *Let $d \in \mathbb{R}$ and Y a centered weakly stationary process with spectral measure ν. Suppose that*

$$\int_{\mathbb{T}} |\iota_d|^2 \, \mathrm{d}\nu < \infty \ . \tag{8.7}$$

Then $\widehat{\mathrm{F}}_{\iota_d}(Y)$ is called the d-order integrated process of Y or the $(-d)$-order differenced process of Y. $\widehat{\mathrm{F}}_{\iota_d}$ is called the integrating operator *of order d or* differencing operator *of order $-d$.*

To explicit the fact that d is not integer valued, one sometimes adds the term *fractionally.* The integer case enjoys specific properties. Note that, if $d \in \mathbb{Z}_-$,

$$\widehat{\mathrm{F}}_{\iota_{-d}} = (\mathbf{1} - \mathrm{B})^{-d} \ .$$

By extension, one often finds the notation $(\mathbf{1}-\mathrm{B})^{-d}$ for all $d \in \mathbb{R}$ in the econometric literature.

Remark 8.3.1. *Note that if $d > 0$, since $|\iota(0)| = \infty$, condition (8.7) implies $\nu(\{0\}) = 0$. In particular, we have that, for all $d \in \mathbb{N}$, if Y a centered weakly stationary process with spectral measure ν satisfying (8.7), we have*

$$\widehat{\mathrm{F}}_{\iota_{-d}} \circ \widehat{\mathrm{F}}_{\iota_d}(Y) = (\mathbf{1} - \mathrm{B})^{d} \circ \widehat{\mathrm{F}}_{\iota_d}(Y) = Y \ .$$

Here we first integrate of order d and differentiate of order d, with $d \geq 0$. If we first differentiate and then integrate, it may no longer be true, since the difference remove any mass at zero frequency, which is not recovered by integrating. More precisely, we find that, for all $d \in \mathbb{Z}_-^*$, if Y a centered weakly stationary process with spectral measure ν satisfying (8.7), we have

$$\widehat{\mathrm{F}}_{\iota_{-d}} \circ \widehat{\mathrm{F}}_{\iota_d}(Y) = \widehat{\mathrm{F}}_{\iota_{-d}} \circ (\mathbf{1} - \mathrm{B})^{-d}(Y) = Y'$$

where

$$Y' = Y - \hat{Y}(\{0\}) \ .$$

See Exercise 8.5.

8.4 Stationary increments processes

For non-positive integers d, since $(\mathbf{1}-\mathrm{B})^{-d}$ is a FIR filter it can be applied to any trajectory of $\mathbb{C}^{\mathbb{Z}}$, hence to any process including non-weakly stationary ones. If d is positive, the definition of $\widehat{\mathrm{F}}_{\iota_{-d}}$ as a mapping on $\mathbb{C}^{\mathbb{Z}}$ is not obvious. For instance, one could ask whether one can defined an operator $(\mathbf{1} - \mathrm{B})^{-1}$ as an inverse of the operator $(\mathbf{1} - \mathrm{B})$ and then $(\mathbf{1} - \mathrm{B})^{-d}$ with d positive integer by iterating d times. We already defined the random walk in Example 8.1.1 which can be seen as process defined on $\mathbb{N}^*$ whose increments are a white noise. Let us extend this to a process indexed by $\mathbb{Z}$.

Example 8.4.1 (Random walk (symmetrized, iterated)). *The random walk of Example 8.1.1 satisfies, for all $t \in \mathbb{N}^*$, setting $X_0 = 0$,*

$$\epsilon_t = X_t - X_{t-1} = [(\mathbf{1} - \mathrm{B})X]_t \ .$$

Moreover, to obtain this relation over $t \in \mathbb{Z}$, the definition of X can be symmetrized by defining, for all $t \in \mathbb{N}_-$,

$$X_t = \sum_{t<k\leq 0} (-\epsilon_k) \ ,$$

Then we have, for all $t \in \mathbb{Z}$,

$$\epsilon_t = X_t - X_{t-1} = [(\mathbf{1} - \mathrm{B})X]_t$$

In fact it is eay to see that X so defined corresponds to the unique process such that $(\mathbf{1}-\mathrm{B})X = \epsilon$ and $X_0 = 0$. Thus this construction can be iterated in order to define $X^{(k)}$ iteratively on $k \geq 1$ such that

$$(\mathbf{1} - \mathrm{B})^k X^{(k)} = \epsilon \quad \textit{and} \quad X_0^{(k)} = 0 \ .$$

In Example 8.4.1, we defined $X = (\epsilon_t)_{t\in\mathbb{Z}}$ such that $(\mathbf{1}-\mathrm{B})X = \epsilon$, hence indicating that we can define an inverse operator to $(\mathbf{1} - \mathrm{B})$ on $\mathbb{C}^{\mathbb{Z}}$. More precisely, we do not use any property on ϵ in the construction, thus it shows that $(\mathbf{1} - \mathrm{B})$ spans $\mathbb{C}^{\mathbb{Z}}$. However this is of course an arbitrary way to do it. As seen in Exercise 8.6, $(\mathbf{1} - \mathrm{B})$ has a null space of dimension 1 on $\mathbb{C}^{\mathbb{Z}}$, hence is not one-to-one. More precisely, since the null space is the space of constant sequences, the operator defining X from ϵ in Example 8.4.1, is the following one.

Definition 8.4.1 (Integrator with zero initial value). *Let I_0 be the $\mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}}$ linear operator uniquely defined by*

$$(\mathbf{1} - \mathrm{B}) \circ I_0 = \mathbf{1}$$

and, for all $x \in \mathbb{C}^{\mathbb{Z}}$, $[I_0(x)]_0 = 0$.

As already said, the operator I_0 is explicitly and quite simply defined in Example 8.4.1 to define $X = I_0(\epsilon)$. Unfortunately, I_0 cannot be used to compute the operator $\widehat{\mathrm{F}}_{\iota_1}$ as in Definition 8.3.2, since the latter gives a weakly stationary process, while I_0 cannot give a (weakly) stationary process other than the trivial zero process. We will see in Exercise 8.2 that, in some cases, $\widehat{\mathrm{F}}_{\iota_d}$ can be written as a causal convolution filter on $\mathbb{C}^{\mathbb{Z}}$ and that, in all cases, it is a causal operator.

One often find the operator $(\mathbf{1} - \mathrm{B})^{-d}$ with $d \in \mathbb{R}$ applied to non-weakly stationary processes or without having the condition (8.7) satisfied. Let us explain what is meant by this in these cases.

To alleviate condition (8.7), one can use the notion of processes with *stationary increments*.

Definition 8.4.2 (Stationary increments processes). *Let $k = 0, 1, 2, \ldots$. A process $X = (X_t)_{t\in\mathbb{Z}}$ is said to have k^{th} order (weakly)* stationary increments *if $(\mathbf{1} - \mathrm{B})^k X$ is (weakly) stationary.*

A weakly stationary increments process $(X_t)_{t\in\mathbb{Z}}$ has a minimal order k of increments that makes it weakly stationary. That is, X is said to be a stationary increments process of minimal order $k \in \mathbb{N}$ if $(\mathbf{1} - \mathrm{B})^k X$ is weakly stationary and, if $k \geq 1$, $Y := (\mathbf{1} - \mathrm{B})^{k-1} X$ is not weakly stationary. The following definition is useful in order to deal with stationary increments processes.

Definition 8.4.3 (Generalized spectral measure)**.** *Let X be a weakly stationary increments process of minimal order $k \in \mathbb{N}$. The* generalized spectral measure *ν of X is defined as follows.*

(i) If $k = 0$, X is weakly stationary and ν is defined as its spectral measure.

(ii) If $k \in \mathbb{N}^*$, ν is defined as the unique measure such that the spectral measure of $(\mathbf{1}-\mathrm{B})^k X$ has density $|\iota_{-k}|^2$ with respect to ν and $\nu(\{0\}) = 0$.

Observe that the minimal order k cannot be recovered from the generalized spectral measure ν as in the following example.

Example 8.4.2 (Trend-stationary time series)**.** *Let Y be a centered weakly stationary process, then the process X defined by adding a* linear trend

$$X_t = Y_t + \beta_0 + \beta_1 t, \qquad t \in \mathbb{Z}\,,$$

has generalized spectral measure equal to the spectral measure of Y but is of minimal order $k = 1$.

Since this order has a strong impact on the way to forecast X, the following definition is widely used in the econometric literature.

Definition 8.4.4 (Integration order)**.** *The* integration order *of a generalized spectral measure ν is defined by*

$$k = \inf\left\{\ell \in \mathbb{Z} \,:\, \int_{\mathbb{T}} |\iota_{-\ell}|^2 \,\mathrm{d}\nu < \infty\right\}\,. \tag{8.8}$$

If the set in the inf is $\mathbb{Z}$ then $k = -\infty$ by convention. A weakly stationary increments process X admits a generalized spectral measure of integration order k, then it is said to be of integration order k*, in short X is I(k). It is morever said to be of* minimal trend degree *if it has stationary increments of minimal order $k_+ = \max(k, 0)$, that is, if one of the two following assertions holds.*

(i) If $k \in \mathbb{N}^*$, X has stationary increments of minimal order equal to the integration order of its generalized spectral measure.

(ii) If $k \leq 0$, X is weakly stationary with spectral measure of integration order k.

Let us examine a simple example.

Example 8.4.3 (Random walk (symmetrized), continued)**.** *Let $k \geq 1$ and consider $X^{(k)}$ as in Example 8.4.1. If $\epsilon = (\epsilon_t)_{t\in\mathbb{Z}}$ is a centered white noise with variance σ^2, then $X^{(k)}$ is an I(k) process with generalized spectral density $|\iota_k|^2$ and it has minimal trend degree.*

We already mentioned in Example 8.4.2 that a process X which has stationary increments with minimal order $k \geq 1$ might not be an I(k) process, hence may not have a minimal trend degree. However it is always equal to an I(k) process with minimal trend degree *up to an additive polynomial process*, as shown in the following result.

Theorem 8.4.1. *Let X be a stationary increments process of minimal order $k \geq 1$ and of integration order $\ell \in \mathbb{Z}$. Then $k \geq \ell$ and there exists a centered I(ℓ) process X' with minimal trend degree such that $X - X'$ is a polynomial process of degree at most k.*

Proof. See Exercise 8.7. □

This result for instance applies to the trend-stationary process in Example 8.4.2 since in this example, X has stationary increments of minimal order 1 and is equal to a weakly stationary centered I(0) process Y (assuming Y to be such) up to polynomial of degree 1.

Example 8.4.4 (Random walk with drift)**.** *A very popular model for log stock prices is the random walk with drift. It is often defined as a process $X = (X_t)_{t\in\mathbb{Z}}$ satisfying the equation*

$$X_t = X_{t-1} + \delta + \epsilon_t \ ,$$

where $\epsilon = (\epsilon_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$. As the random walk without drift of Example 8.4.1 or the trend-stationary time series of Example 8.4.2, the random walk with drift has stationary increments of minimal order 1. It is an $I(1)$ process (hence with minimal trend degree) and is equal to a random walk without drift (hence a centered $I(1)$ process with minimal trend degree) up to a polynomial of degree 1.

The main interest of the above definitions is to be able to apply the integration operator at any order.

Theorem 8.4.2. *Let $k \in \mathbb{Z}$ and X be a centered $I(k)$ process with minimal trend degree and let ν be its generalized spectral measure. Assume that $\nu(\{0\}) = 0$. Let $d \in \mathbb{R}$ and let*

$$\ell := \min\left\{ j \in \mathbb{Z} \ : \ \int_{\mathbb{T}} |\iota_{d-j}|^2 \ \mathrm{d}\nu < \infty \right\} \ ,$$

which belongs to $[k + d, k + d + 1)$. Then the process

$$I^{(d)}(X) := I_0^{\ell_+} \circ \widehat{\mathrm{F}}_{\iota_{d+k_+-\ell_+}} \circ (\mathbf{1} - \mathrm{B})^{k_+}(X) \tag{8.9}$$

is well defined and it is an $I(\ell)$ process with minimal trend degree whose generalized spectral measure ν' has density $|\iota_d(\lambda)|^2$ with respect to ν. It is called the generalized integrated process *of order d of X.*

Proof. First note that since we assumed that $\nu(\{0\}) = 0$, we have

$$\int_{\mathbb{T}} |\iota_{d-j}|^2 \ \mathrm{d}\nu = \int_{\mathbb{T}} |\iota_{-j}|^2 \ \mathrm{d}\nu' \ .$$

Hence ℓ is the integration order of ν'. Thus, we only have to show that $I^{(d)}(X)$ is well defined, is weakly stationary with spectral measure ν' if $\ell \leq 0$ and has weakly stationary increments with minimal order ℓ and generalized spectral measure ν' and if $\ell \geq 1$.

Observe that $(\mathbf{1} - \mathrm{B})^{k_+}(X)$ is centered weakly stationary with spectral measure

$$\left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{2k_+} \nu(\mathrm{d}\lambda) \ ,$$

and we have

$$\int_{\mathbb{T}} \left|\iota_{d+k_+-\ell_+}(\lambda)\right|^2 \left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{2k_+} \ \nu(\mathrm{d}\lambda) \leq \int_{\mathbb{T}} \left|\iota_{d-\ell_+}\right|^2 \ \nu(\mathrm{d}\lambda) < \infty \ ,$$

by definition of ℓ. Hence

$$Y := \widehat{\mathrm{F}}_{\iota_{d+k_+-\ell_+}} \circ (\mathbf{1}-\mathrm{B})^{k_+}(X)$$

is a well defined centered weakly stationary process and, so $I^{(d)}(X)$ in (8.9) is well defined and we have that $(\mathbf{1}-\mathrm{B})^{\ell_+}\left(I^{(d)}(X)\right)$ is this weakly stationary process Y and thus has spectral measure

$$\left|\iota_{d-\ell_+}(\lambda)\right|^2 \, \nu(\mathrm{d}\lambda) = \left|\iota_{-\ell_+}(\lambda)\right|^2 \, \nu'(\mathrm{d}\lambda) \, .$$

If $\ell \leq 0$, we get that $I^{(d)}(X)$ is weakly stationary with spectral measure ν'.

Suppose now that $\ell \geq 1$; it remains to show that $(\mathbf{1}-\mathrm{B})^{\ell-1}\left(I^{(d)}(X)\right)$ is not stationary. Note that, in this case, we have

$$(\mathbf{1}-\mathrm{B})^{\ell-1}\left(I^{(d)}(X)\right) = I_0(Y) \, ,$$

which is stationary only if Y is the zero process, which would imply ν to be zero. But this is not possible if $\ell \geq 1$. □

Example 8.4.5 (ARFIMA(0,*d*,0) process)**.** *Take $\epsilon \sim \mathrm{WN}(0,\sigma^2)$ for some $\sigma^2 > 0$ and let $d \in \mathbb{R}$. The process $I^{(d)}(\epsilon)$ is a stationary increments process with generalized spectral density*

$$f(\lambda) = \left|1-\mathrm{e}^{-\mathrm{i}\lambda}\right|^{-2d} \, .$$

It is weakly stationary if and only if $d < 1/2$. This process is called an ARFIMA(0,d,0) process. We will see in Section 8.5 the genral class of ARFIMA processes.

8.5 AR(F)IMA processes

Because $\mathbf{1}-\mathrm{B}$ is a polynomial of B with root 1, in the econometric literature, an $I(1)$ process is often called *unit root* non-stationary. This is because these processes are essentially seen as extension of the class of ARMA processes, where unit roots are allowed in the AR part. More generally, the following definition also takes into account possible unit roots on the MA part and exclude any other roots with modulus 1 both on the AR and MA parts.

Definition 8.5.1 (ARIMA processes)**.** *Let $d \in \mathbb{Z}$ and $p,q \in \mathbb{N}$. A centered process $X = (X_t)_{t\in\mathbb{Z}}$ is said to be* autoregressive integrated moving averages *or order (p,d,q) (ARIMA(p,d,q)) process if the following assertions hold, depending on the sign of d.*

1. *In the case where $d \geq 0$, $(\mathbf{1}-\mathrm{B})^d(X)$ is a canonical ARMA(p,q) process.*
2. *In the case where $d < 0$, X is an ARMA$(p, d+q)$ process such that $\widehat{\mathrm{F}}_{\iota_{-d}}(X)$ is well defined and admits a canonical ARMA(p,q) representation.*

Hence X is ARIMA(p,d,q) if and only if there exist two polynomials of degree q and p

$$\Theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k \quad \text{and} \quad \Phi(z) = 1 - \sum_{k=1}^{q} \phi_k z^k$$

which do not vanish on the unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$, and $Z \sim \mathrm{WN}(0,\sigma^2)$ such that X is solution to

$$\begin{cases} (\mathbf{1}-\mathrm{B})^d \circ \Phi(\mathrm{B})(X) = \Theta(\mathrm{B})(Z) & \text{if } d \geq 0, \\ \Phi(\mathrm{B})(X) = (\mathbf{1}-\mathrm{B})^{-d} \circ \Theta(\mathrm{B})(Z) & \text{otherwise.} \end{cases}$$

We then easily get the following result.

Theorem 8.5.1. *Let $d \in \mathbb{Z}$, and $p, q \in \mathbb{N}$. A centered process X is ARIMA(p,d,q) if and only if it is an $I(d)$ process with minimal trend degree and generalized spectral density of the form*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{-2d} \left|\frac{\Theta(\mathrm{e}^{-\mathrm{i}\lambda})}{\Phi(\mathrm{e}^{-\mathrm{i}\lambda})}\right|^2 ,$$

where $\sigma^2 > 0$, and Φ and Θ are polynomials of degrees p and q which do not vanish on the unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ and take value 1 at 0.

ARIMA processes are widely used for forecasting time series. Note that, in this definition, X is uniquely defined by Z, Φ and Θ if and only if $d \leq 0$. If $d > 0$, only $(\mathbf{1}-\mathrm{B})^d(X)$ is uniquely defined, hence X is defined up to a polynomial process of degree $d-1$. Nevertheless, for any $d \in \mathbb{Z}$, when one refers to *the* ARIMA(p,d,q) process with innovation Z, AR polynomial Φ and MA polynomial Θ, one usually refers to the process

$$X = I^{(d)} \circ \Phi(\mathrm{B})^{-1} \circ \Theta(\mathrm{B})(Z) ,$$

where $I^{(d)}$ is defined as in Theorem 8.4.2. We can apply this theorem here since $\Phi(\mathrm{B})^{-1} \circ \Theta(\mathrm{B})(Z)$ is a canonical ARMA hence an I(0) process. It thus follows that this definition extends to any $d \in \mathbb{R}$, giving raise to the following definition.

Definition 8.5.2 (ARFIMA processes)**.** *Let $d \in \mathbb{R}$ and $p, q \in \mathbb{N}$. A centered process $X = (X_t)_{t\in\mathbb{Z}}$ is said to be* autoregressive fractionally integrated moving averages *or order (p,d,q) (ARFIMA(p,d,q)) process if there exists $k \in \mathbb{Z}$ such that X is an $I(k)$ process with minimal trend degree and its generalized spectral density is of the form*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - \mathrm{e}^{-\mathrm{i}\lambda}\right|^{-2d} \left|\frac{\Theta(\mathrm{e}^{-\mathrm{i}\lambda})}{\Phi(\mathrm{e}^{-\mathrm{i}\lambda})}\right|^2 ,$$

where $\sigma^2 > 0$, and Φ and Θ are polynomials of degrees p and q which do not vanish on the closed unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ and take value 1 at 0.

Let us compute the integer k appearing in this definition. Since the given generalized spectral density is continuous out of the origin and behaves when $\lambda \to 0$ as

$$f(\lambda) \sim \frac{\sigma^2}{2\pi} \left|\frac{\Theta(1)}{\Phi(1)}\right|^2 |\lambda|^{-2d} ,$$

with Θ and Φ non-vanishing on the closed unit disk, we have

$$k = \min\{\ell \in \mathbb{Z} : 2\ell - 2d > -1\} = \min\{\ell \in \mathbb{Z} : \ell > d - 1/2\} .$$

(k is the smallest integer strictly larger than $d-1/2$.) Then, applying Theorem 8.4.2 for any real d, we get the following.

Theorem 8.5.2. *Let $d \in \mathbb{R}$, and $p, q \in \mathbb{N}$. Suppose that*

$$X = I^{(d)}(Y) ,$$

where Y admits a canonical ARMA(p,q) representation and $I^{(d)}$ is defined as in Theorem 8.4.2. Then X is ARFIMA(p,d,q) and satisfies

$$[(\mathbf{1} - \mathrm{B})^j X](0) = 0 \qquad \text{for all } 0 \leq j \leq d - 1/2 . \tag{8.10}$$

Proof. The result is a straightforward application of Theorem 8.4.2 and Definition 8.5.2. □

Note that Condition (8.10) is empty if $d < 1/2$. If $d \geq 1/2$ this condition characterizes the process X among all ARFIMA(p, d, q) processes such that $Y = I^{(-d)}(X)$ and these processes are equal to X up to a polynomial of degree at most $k - 1$.

8.6 Functional limit of partial sums

To decide whether a process is an I(0) process, one can investigate the statistical behavior of its partial sums statistics.

Definition 8.6.1 (Partial sums)**.** *The* partial sums *of a sample $X_1, \ldots, X_n$ is defined as the $\mathbb{R}^{[0,1]}$-valued statistic $S_n = (S_n(t))_{t\in[0,1]}$ defined by*

$$S_n(t) = \sum_{k=1}^{[tn]} X_k \quad \textit{for all} \quad t \in [0, 1] \ ,$$

where $[x]$ denotes the entire part of x and with the convention $\sum_{k=1}^{0} \cdots = 0$.

The partial sums is functional valued statistic, but it belongs to a very restricted class of functions, namely the *piecewise constant* functions, since it is constant on intervals $[k/n, (k + 1)/n)$. Note also that by definition of the entire part, S_n is a *cadlag* function, that is, a function which is right-continuous and admits left limits at every points.

Let us examine the behavior of the statistic S_n as $n \to \infty$, depending on the order of the process X. For a fixed $t \in (0, 1]$, we have

$$S_n(t) = [nt]\hat{\mu}_{[nt]} \ ,$$

where $\hat{\mu}_n$ is the usual empirical mean from the sample $X_1, \ldots, X_n$. Hence for a "standard" weakly stationary process X with mean μ and spectral density f, we expect that

$$S_n(t) \sim t\mu \, n \quad \text{a.s.}$$

However this behavior is merely related to the value of the mean μ, rather than the covariance structure of the process. To get rid of the mean, we consider the *bridged* partial sums $P_n = (P_n(t))_{t\in[0,1]}$ defined by

$$P_n(t) = S_n(t) - tS_n(1) \quad \text{for all} \quad t \in [0, 1] \ . \tag{8.11}$$

Lemma 8.6.1. *Suppose that the sample $X_1, \ldots, X_n$ has a constant mean μ. Let S_n and P_n be its partial sums and bridged partial sums, and denote their centered versions by*

$$\overline{S}_n(t) = \sum_{k=1}^{[tn]} (X_k - \mu) \quad \textit{and} \quad \overline{P}_n(t) = \overline{S}_n(t) - t\overline{S}_n(1) \ .$$

then we have, for all $t \in [0, 1]$,

$$\left|P_n(t) - \overline{P}_n(t)\right| \leq |\mu| \ .$$

Proof. We have that

$$S_n(t) - tS_n(1) = \sum_{k=1}^{[tn]}(X_k - \mu) - t\sum_{k=1}^{n}(X_k - \mu) + (tn - [tn])\mu \ .$$

Hence the result. □

Hence by using P_n instead of S_n, the effect of a non-zero constant mean is bounded by a constant not depending on n.

Next, we examine the behavior of the statistics S_n and P_n as $n \to \infty$ in two different cases:

(A-1) X is an I(k) process with $k \leq -1$.

(A-2) X is a weakly stationary process and admits a density which is strictly positive and continuous at zero.

Observe that Case (A-2) implies that X is $I(0)$. (Actually by Exercise 8.3, this case contains all I(0) processes with *smooth* enough spectral density at the origin.) We first state some simple L^2 behavior and then, under additional assumptions, satisfied for instance by ARIMA processes, more elaborated asymptotic results.

Lemma 8.6.2. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a real valued weakly stationary process with mean μ. Let S_n and P_n be the partial sums and bridged partial sums obtained from the sample $X_1, \ldots, X_n$. The following assertions hold.*

(i) If (A-1) holds, then

$$\limsup_{n\to\infty} \sup_{t\in[0,1]} \mathrm{Var}\,(S_n(t)) < \infty \ .$$

(ii) If (A-2) holds, then, for all $t \in (0, 1]$, as $n \to \infty$,

$$\liminf_{n\to\infty} n^{-1}\ \mathrm{Var}\,(S_n(t)) > 0 \ .$$

Proof. Suppose first that (A-1) holds. Then there exists a weakly stationary process Z such that $X = (\mathbf{1} - \mathrm{B})Z$. It follows that, for all $t \in [0, 1]$,

$$S_n(t) = Z_{[nt]} - Z_0 \ .$$

Assertion (i) follows.

Suppose now that (A-2) holds and denote by f the spectral density of X.

Note now that, we have

$$(2\pi n)^{-1}\,\mathrm{Var}\,(S_n(t)) = \int_{\mathbb{T}} J_{[nt]}(\lambda)\, f(\lambda)\ \mathrm{d}\lambda \ ,$$

where J_n is the Fejér kernel defined in Exercise A.2,

$$J_n(t) = \frac{1}{2\pi}\sum_{k=-n+1}^{n-1}(1 - |k|/n)\mathrm{e}^{\mathrm{i}kt} = \frac{1}{2\pi n}\left|\sum_{j=0}^{n-1}\mathrm{e}^{\mathrm{i}jt}\right|^2 \ .$$

Using that $\int_{\mathbb{T}} J_n = 1$ and that, for all $\epsilon \in (0, \pi)$,

$$\sup_{n\geq 1} \sup_{\epsilon \leq |\lambda| \leq \pi} n J_n(\lambda) < \infty \ ,$$

we get that

$$\lim_{n\to\infty} (2\pi n)^{-1} \operatorname{Var}\left(S_n(t)\right) = t\, f(0) \ .$$

Hence the result. □

Lemma 8.6.2 is an indication of the fact that S_n has a different asymptotic behavior depending of the order of integration of X, and that this is related to the value of the spectral density at the origin, when there is one. This should not surprise us, since we have

$$\frac{1}{\sqrt{n}} \overline{S}_n(t) = \frac{\sqrt{[tn]}}{\sqrt{n}} \, \frac{1}{\sqrt{[tn]}} \sum_{k=1}^{[tn]} (X_k - \mu) \Longrightarrow \mathcal{N}(0, 2\pi\, t\, f(0)) \ ,$$

where the latter weak convergence holds under Assumption 4.3.1 and Assumption 4.4.1 by Theorem 4.4.2 and Slutsky's lemma. This iresult is not completly satisfactory as it requires the (usually unknown) mean μ. By Lemma 8.6.1, it seems natural to replace S_n by P_n as a possible statistic that should not need centering. Indeed the previous convergence immediately gives that, for all $t \in (0, 1]$,

$$P_n(t) = S_n(t) - t S_n(1) = O_p(\sqrt{n}) \ .$$

However to get the weak convergence of $n^{-1/2} P_n$, we need the *joint convergence* of $n^{-1/2}(\overline{S}_n(t), \overline{S}_n(1))$. More generally, we obviously get the *fidi* convergence of $\left(n^{-1/2} P_n(t)\right)_{t\in[0,1]}$ from that of $\left(n^{-1/2} \overline{S}_n(t)\right)_{t\in[0,1]}$. Concerning the latter we have the following.

Proposition 8.6.3. *Suppose that Assumptions 4.3.1 and 4.4.1 hold. Then we have*

$$\frac{1}{\sqrt{n}} \, \overline{S}_n \stackrel{\text{fidi}}{\Longrightarrow} \sqrt{2\pi\, f(0)}\, B \ ,$$

where $(B(t))_{t\in[0,1]}$ *is a centered Gaussian process with* stationary and independent *increments and such that* $B(0) = 0$ $\mathbb{P}$-a.s. *and* $\operatorname{Var}(B(1)) = 1$.

Proof (sketch). The proof essentially elaborates around the proof of Theorem 4.4.2 and we only sketch to explain how the Brownian motion appears in the limit. Let $p \in \mathbb{N}^*$ and $0 < t_1 < t_2 < \cdots < t_p \leq 1$.

$$\begin{array}{cccc}
\underbrace{X_1, X_2, \ldots, X_{[t_1 n]}}_{\overline{S}_n(t_1)}, & \underbrace{X_{[t_1 n]+1}, \ldots, X_{[t_2 n]}}_{\overline{S}_n(t_2) - \overline{S}_n(t_1)}, & \ldots, & \underbrace{X_{[t_{p-1} n]+1}, \ldots, X_{[t_p n]}}_{\overline{S}_n(t_p) - \overline{S}_n(t_{p-1})} \\
\shortparallel & & & \\
[t_1 n](\hat{\mu}_{[t_1 n]} - \mu) & \ldots & \ldots & \ldots
\end{array}$$

Elaborating on the proof of the asymptotic normality of $\hat{\mu}_n$, it may be shown that these increments are *asymptotically normal and independent*. More precisely, we get that

$$\frac{1}{\sqrt{n}} \begin{bmatrix} \overline{S}_n(t_1) \\ \overline{S}_n(t_2) - \overline{S}_n(t_1) \\ \vdots \\ \overline{S}_n(t_p) - \overline{S}_n(t_{p-1}) \end{bmatrix} \Longrightarrow \mathcal{N}\left(0, 2\pi\, f(0) \begin{bmatrix} t_1 & 0 & \ldots & 0 \\ 0 & t_2 - t_1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & t_p - t_{p-1} \end{bmatrix}\right) \ .$$

The fidi convergence follows. □

The existence of the process B in Proposition 8.6.3 is a byproduct of the proof.

Definition 8.6.2 (Standard Brownian motion)**.** *The centered Gaussian process* $(B(t))_{t\in[0,1]}$ *with stationary and independent increments and such that* $B(0) = 0$ $\mathbb{P}$-a.s. *and* $\mathrm{Var}\,(B(1)) = 1$ *is called the* standard Brownian motion *defined on* $[0, 1]$.

The fidi convergence allows to derive the limit of

$$g\left(\frac{1}{\sqrt{n}}\,\overline{S}_n(t_i),\, i = 1, \ldots, p\right)$$

for all $0 \le t_1 < t_2 < \cdots < t_p \le 1$ and every *continuous* $g : \mathbb{R}^p \to \mathbb{R}$. In particular, using Lemma 8.6.1, we get the following.

Corollary 8.6.4. *Suppose that Assumptions 4.3.1 and 4.4.1 hold. Then we have*

$$\frac{1}{\sqrt{n}}\,P_n \overset{\text{fidi}}{\Longrightarrow} \sqrt{2\pi\, f(0)}\; P$$

where $(P(t))_{t\in[0,1]}$ *is the* standard Brownian bridge*,*

where we have used the following.

Definition 8.6.3. *The process defined on* $[0, 1]$ *by*

$$P(t) = B(t) - tB(1)\,, \qquad t \in [0, 1]\,,$$

where $(B(t))_{t\in[0,1]}$ *is the standard Brownian motion, is called the* standard Brownian bridge*.*

However, we cannot deduce anything about

$$\sup_{t\in[0,1]} \overline{S}_n(t), \quad \int_0^1 \overline{S}_n(t)\,\mathrm{d}t, \quad \ldots$$

from the fidi convergence. In fact, a much more precise result can be obtained than the fidi convergence. We will only provide the basic necessary facts concerning this extension, without proofs. We refer to Billingsley [1999b] for the reader who would like to learn more on this topic (starting with the Donsker theorem). one can show that the Brownian bridge admits a continuous version, that is, we can defined a process B valued in the space of $[0, 1] \to \mathbb{R}$ continuous functions endowed with the Borel σ-field associated to the sup norm such that the fidi distribution of B is that of the Brownian motion. Unfortunately S_n is not continuous and neither $\bar{S}_n$, nor P_n are. However, as we already mentioned, they are in the class of $[0, 1] \to \mathbb{R}$ cadlag functions, that we denote by $\mathbb{D}$. It turns out that $\mathbb{D}$ can be endowed with a metric, called the J_1 Skorohod metric (see Billingsley [1999b]), which makes $\mathbb{D}$ a separable complete metric space. In the following, we admit that the standard Brownian motion defined on $[0, 1]$ admits a modification that makes it having continuous trajectories, hence valued in $\mathbb{D}$. We will continue to call this process the standard Brownian motion, and of course, the same is true about the standard Brownian bridge. We will need the following result, here stated without proof, see e.g. Billingsley [1999b] for the i.i.d. case.

Theorem 8.6.5 (Donsker theorem, Invariance principle)**.** *Let $(B(t))_{t\in[0,1]}$ denote the standard Brownian motion and $\mathbb{D}$ be the class of $[0,1] \to \mathbb{R}$ cadlag functions endowed with the Borel σ-field associated to the sup norm. The following assertions hold. Let $Z \sim \mathrm{IID}(0,1)$. Then the following* invariance principle *holds.*

$$\frac{1}{\sqrt{n}} S_n^Z \Longrightarrow B \qquad \text{in } \mathbb{D}, \tag{8.12}$$

where S_n^Z denotes the partial sums statistic associated to the process Z.

This theorem can be extended to non-iid time series, for instance for X defined as a linear process as in Assumption 4.3.1 with additional assumptions on the noise Z, see Davydov [1970] or Peligrad and Utev [2006]. Then, as in Theorem 4.4.2, the asymptotic variance has to be adapted into

$$\frac{1}{\sqrt{n}} \bar{S}_n \Longrightarrow \sqrt{2\pi\, f(0)}\, B \qquad \text{in } \mathbb{D}, \tag{8.13}$$

where $\bar{S}_n$ denotes the centered partial sums statistic associated to the process X.

Of course the following result immediately follows.

Corollary 8.6.6. *Suppose that (8.13) holds. Then we have*

$$\frac{1}{\sqrt{n}} P_n \Longrightarrow \sqrt{2\pi\, f(0)}\, P \qquad \text{in } \mathbb{D}.$$

where $(P_n(t))_{t\in[0,1]}$ is defined in (8.11) and $(P(t))_{t\in[0,1]}$ is the standard Brownian bridge.

8.7 Unit root test

The results of Section 8.6 show that the partial sum statistics should be able to help distinguishing between a zero order and a negative order of integration. However it is more classical in the econometric literature to consider the problem of distinguishing between order 1 of integration and zero order in a special parametric case, namely, where the zero order corresponds to an AR(1) process and the the order 1 to a pure random walk, leading to the *Dickey-Fuller* test (DF test). Extensions to the random walk with drift or more sophisticated time trend can be considered as well as more elaborated models, in particular those yielding the so called *augmented Dickey-Fuller* test (ADF test).

Hence we consider the problem is to construct a test from a sample $X_1, \ldots, X_n$ to distinguish between the hypotheses

(H-0) X is a pure random walk ,

$$X_t = \sum_{k=1}^{t} \epsilon_k, \quad \text{where} \quad \epsilon \sim \mathrm{IID}(0, \sigma^2)\ .$$

(H-1) X is a stationary AR(1) process,

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \text{where} \quad |\phi| < 1,\ \epsilon \sim \mathrm{IID}(0, \sigma^2)\ .$$

The idea behind the construction of the DF test is to these two hypotheses as two subsets of the linear model

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \text{for all} \quad t = 2, \dots, n \ ,$$

with $\phi = 1$ under (H-0) and $|\phi| < 1$ under (H-1). We thus proceed with the usual approach for linear models, introducing the least square estimator

$$\begin{aligned} \hat{\phi}_n &= \operatorname*{argmin}_{a \in \mathbb{R}} \sum_{t=2}^{n} (X_t - a\, X_{t-1})^2 \\ &= \frac{\sum_{t=2}^{n} X_{t-1} X_t}{\sum_{t=2}^{n} X_{t-1}^2} \\ &= \phi + \frac{\sum_{t=2}^{n} X_{t-1} \epsilon_t}{\sum_{t=2}^{n} X_{t-1}^2} \ . \end{aligned} \tag{8.14}$$

Then, we get that, under (H-1),

$$\hat{\phi}_n = \phi + O_p(n^{-1/2}) \ .$$

Under (H-0), in turn, due to the non-stationarity of X, the behavior of $\hat{\phi}_n$ is much more involved. Let

$$S_n(t) = \sum_{k=1}^{[tn]} \epsilon_k, \quad t \in [0, 1] \ .$$

denote the partial sum statistic associated to ϵ. Then we have, under (H-0),

$$\frac{1}{n} \sum_{t=2}^{n} X_{t-1}^2 = \sum_{t=2}^{n} \int_{(t-1)/n}^{t/n} S_n^2(u)\, \mathrm{d}u = \int_0^1 S_n^2(u)\, \mathrm{d}u \ . \tag{8.15}$$

On the other hand, under (H-0),

$$2 \sum_{t=2}^{n} X_{t-1} \epsilon_t = 2 \sum_{1 \le s < t \le n} \epsilon_s \epsilon_t = \left(\sum_{t=1}^{n} \epsilon_t \right)^2 - \sum_{t=1}^{n} \epsilon_t^2 \ ,$$

and thus,

$$\frac{1}{n} \sum_{t=2}^{n} X_{t-1} \epsilon_t = \frac{1}{2} (n^{-1/2} S_n(1))^2 - \frac{1}{2n} \sum_{t=1}^{n} \epsilon_t^2 \ . \tag{8.16}$$

Since $\epsilon \sim \mathrm{IID}(0, \sigma^2)$, we have

$$\frac{1}{2n} \sum_{t=1}^{n} \epsilon_t^2 \xrightarrow{\text{a.s.}} \frac{\sigma^2}{2} \ ,$$

and the Donsker theorem gives that

$$\frac{1}{\sqrt{n}}\, S_n \Longrightarrow \sigma\, B \quad \text{in } \mathbb{D} \ .$$

With (8.14), (8.15) and (8.16), we finally get that, under (H-0),

$$n \left(\hat{\phi}_n - 1 \right) \Longrightarrow \frac{B^2(1) - 1}{2 \displaystyle\int_0^1 B^2(u)\, \mathrm{d}u} \ ,$$

whose distribution is called the *Dickey-Fuller distribution.*

Define the test statistic

$$T_n = n \left| \hat{\phi}_n - 1 \right| ,$$

so that, under (H-0), T_n has a known asymptotic distribution,

$$T_n \Longrightarrow \frac{\left| B^2(1) - 1 \right|}{2 \displaystyle\int_0^1 B^2(u) \, \mathrm{d}u} ,$$

while, under (H-1),

$$T_n = n|\phi - 1| + O_p(\sqrt{n}) .$$

The quantile and distribution functions of the Dickey-Fuller distribution are available in standard statistical or time series software. In such a software, the Dickey-Fuller (DF) test is usually directly implemented in a more general fashion than above, namely by adding a drift (as in Example 8.4.4) or an additional trend, so that the hypotheses typically become

(H-0) X is a random walk with trend ,

$$X_t = \delta + \alpha\, t + \sum_{k=1}^{t} \epsilon_k, \quad \text{where} \quad \delta, \alpha \in \mathbb{R} \ , \ \epsilon \sim \mathrm{IID}(0, \sigma^2) \ .$$

(H-1) X is a stationary AR(1) process with a linear trend,

$$X_t = \delta + \alpha\, t + \phi X_{t-1} + \epsilon_t, \quad \text{where} \quad |\phi| < 1, \delta, \alpha \in \mathbb{R} \ , \ \epsilon \sim \mathrm{IID}(0, \sigma^2) \ .$$

A further extension referred to as the Augmented Dickey-Fuller (ADF) test, which builds on the AR model, yielding the hypotheses, for a given $p \in \mathbb{N}^*$,

(H-0) X is obtained by integrating an AR$(p-1)$ with trend ,

$$X_t = \delta + \alpha\, t + \sum_{k=1}^{t} Y_k, \quad \text{where} \quad \delta, \alpha \in \mathbb{R} \ , \ Y \sim \mathrm{AR}(p-1) \ .$$

(H-1) X is a stationary AR(p) process with a linear trend,

$$X_t = \delta + \alpha\, t + \sum_{k=1}^{p} \phi_k X_{t-1} + \epsilon_t, \quad \text{where} \quad \delta, \alpha \in \mathbb{R} \ , \ \epsilon \sim \mathrm{IID}(0, \sigma^2) \ ,$$

and the AR coefficients $\phi_1, \ldots, \phi_p$ satisfy $1 - \sum_{j=1}^{p} \phi_k z^k \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$.

Again, in the econometric literature, these two hypotheses are often rewritten as a general linear model. This is done by introducing the increments of X. Droping the trend terms for simplicity, consider the AR(p) equation

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t \ .$$

One can rewrite this equation as

$$X_t = \tilde{\phi}_1 X_{t-1} + \sum_{k=2}^{p} \tilde{\phi}_k (\Delta X)_{t-k+1} + \epsilon_t \ ,$$

where we set

$$\begin{aligned} \Delta X &= (\mathbf{1} - \mathrm{B}) X \ , \\ \tilde{\phi}_k &= -\left(\phi_k + \cdots + \phi_p\right) \ , \qquad 1 \leq k \leq p \ , \\ \tilde{\phi}_1 &= \phi_1 + \cdots + \phi_p \ . \end{aligned}$$

The presence of a unit root in the original AR equation is then equivalent to $\tilde{\phi}_1 = 1$.

8.8 Exercises

Exercise 8.1. Let α and β be two measurable functions from $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ to $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$. Let X be a centered weakly stationary process with spectral measure ν and $Y = \widehat{\mathrm{F}}_\alpha(X)$.

1. Determine, under a suitable assumption on α and ν, the spectral measure ν' of Y.
2. Let β_n be a sequence of continuous $\mathbb{T} \to \mathbb{C}$ functions. Show that they converge to β in $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu')$ if and only if $\alpha\beta_n$ converge to $\alpha\beta$ in $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$.
3. Show that, for all $t \in \mathbb{Z}$,
$$\int \mathrm{e}^{\mathrm{i}\lambda t} \, \mathrm{d}\hat{Y}(\lambda) = \int \mathrm{e}^{\mathrm{i}\lambda t} \, \alpha(\lambda) \, \mathrm{d}\hat{X}(\lambda) \, .$$
4. Deduce that, under a suitable assumption on β and ν', we have
$$\int \beta(\lambda) \, \mathrm{d}\hat{Y}(\lambda) = \int [\beta\alpha](\lambda) \, \mathrm{d}\hat{X}(\lambda) \, .$$
5. Conclude the proof of Proposition 8.2.3.

Exercise 8.2. Let $d \in \mathbb{R}$ and Y be a centered weakly stationary process with spectral measure ν. Define $c(d) = (c_k(d))_{k\in\mathbb{Z}}$ by
$$c_k(d) = 0 \qquad \text{for all } k = -1, -2, \ldots$$
and
$$(1-z)^{-d} = \sum_{k=0}^{\infty} c_k(d) \, z^k \qquad \text{for all } z \in \mathbb{C} \text{ such that } |z| < 1.$$

1. Compute $c_k(d)$.
2. What is the behavior of $c_k(d)$ as $k \to \infty$ if $d \in \mathbb{Z}_-$?
3. If $d \notin \mathbb{Z}_-$, show that there exists $k_0 \in \mathbb{N}$ such that $(d-1)/k_0 > -1$ and, for all $k \geq k_0$,
$$\ln \frac{c_k(d)}{c_{k_0-1}(d)} = \sum_{j=k_0}^{k} \ln(1 + (d-1)/j) \, .$$
4. Deduce that, if $d \notin \mathbb{Z}_-$, there exists $c > 0$ such that, for all $k \geq k_0$,
$$c^{-1} \, k^{d-1} \leq \frac{c_k(d)}{c_{k_0-1}(d)} \leq c \, k^{d-1} \, .$$

We temporiraly assume that Y has a bounded spectral density.

5. Determine all d's such that
$$\iota_d(\lambda) = \sum_{k\in\mathbb{Z}} c_k(d) \mathrm{e}^{-\mathrm{i}k\lambda}$$
where the convergence holds in $L^2(\mathbb{T})$ endowed with the Lebesgue measure.

6. Show that, if $d < 1/2$, for all $t \in \mathbb{Z}$,

$$\left[\widehat{\mathrm{F}}_{\iota_d}(Y)\right]_t = \sum_{k=0}^{\infty} c_k(d) Y_{t-k} \ ,$$

   where the convergence holds in L^2.

We now assume what is needed to define $\widehat{\mathrm{F}}_{\iota_d}(Y)$, that is that

$$\int |\iota_d|^2 \ \mathrm{d}\nu < \infty \ .$$

7. Show that, as $a \uparrow 1$, $\lambda \mapsto (1 - a\mathrm{e}^{-\mathrm{i}\lambda})^{-d}$ converges to ι_d in $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu)$.

8. Conclude that $\widehat{\mathrm{F}}_{\iota_d}$ is causal, that is, for all $t \in \mathbb{Z}$,

$$\left[\widehat{\mathrm{F}}_{\iota_d}(Y)\right]_t \in \mathcal{H}_t^Y \ .$$

Exercise 8.3. Let X be a real valued weakly stationary process with spectral density f.

1. Show that if f is continuous at zero and $f(0) > 0$ then X is I(0).

2. Reciprocally, show that if f is two times differentiable at zero and X is I(0) then $f(0) > 0$.

Exercise 8.4. Give an example of a non-trivial weakly stationary increments I($-\infty$) process.

Exercise 8.5. Let Y be a centered weakly stationary process with spectral measure ν. Suppose that $\nu(\{0\}) > 0$ and denote ν^* the measure defined by

$$\nu^* = \nu(\cdot \cap (\mathbb{T} \setminus \{0\})) \ .$$

Let $Y' = Y - \hat{Y}(\{0\})$. Show that Y' is weakly stationary and compute its spectral measure.

Exercise 8.6. For any $\ell \geq -1$, we denote by $\mathcal{P}_\ell$ the space of polynomial sequences in $\mathbb{C}^{\mathbb{Z}}$ of degree at most ℓ, with the convention that the polynomial of degree -1 is the zero polynomial.

1. What is the null space of the linear application $\mathbf{1} - \mathrm{B}$ on $\mathbb{C}^{\mathbb{Z}}$?

2. Let $\ell \in \mathbb{N}$. Show that the image of $\mathcal{P}_\ell$ by $\mathbf{1} - \mathrm{B}$ is $\mathcal{P}_{\ell-1}$.

3. Let $\ell \in \mathbb{N}$. Show that $x \in \mathcal{P}_{\ell-1}$ if and only if

$$(\mathbf{1} - \mathrm{B})^\ell x = 0 \ .$$

   [**Hint**: use a recursion on ℓ.]

Exercise 8.7. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a weakly increment stationary process of minimal order $k \geq 1$ and generalized spectral measure ν with integration order $\ell \in \mathbb{Z}$. Define $Y := (\mathbf{1}-\mathrm{B})^k X$ and $Y' = Y - \mu^Y$, where μ^Y is the mean of Y.

1. Show that $k \geq \ell$.

2. Show that $\widehat{\mathrm{F}}_{\iota_{k-\ell}}(Y')$ is well defined.

3. Show that, if $\ell \leq 0$, $\widehat{\mathrm{F}}_{\iota_k}(Y')$ is well defined.

We now define the process X' by

$$X' = \begin{cases} I_0^\ell \circ \widehat{\mathrm{F}}_{\iota_{k-\ell}}(Y') & \text{if } \ell \geq 1 \\ \widehat{\mathrm{F}}_{\iota_k}(Y') & \text{otherwise} , \end{cases}$$

where I_0 is as in Definition 8.3.2.

4. Show that $X - X'$ is a polynomial process of degree at most k.

5. Show that X' is an $I(\ell)$ process with generalized spectral measure ν.

Chapter 9

Cointegration

In this chapter, we consider vector valued (or multivariate) time sries. This type of models have been encountered already, in particular in Chapter 7 and all the models that we describe here can be seen as specific cases of the dynamic linear models introduced therein. However, here, we focus on multivariate time series with a vector AR (VAR) structure and investigate some specific non-stationary behaviors in this context, which we did not look at much in Chapter 7. Nonetheless many algorithms introduced in Chapter 7 for forecasting or computing the Gaussian likelihood can be used in the present context.

Nonstationary behaviors of univariate timeseries are often referred to as *trends*. From a modelling point of view, trends are either deterministic, in the form of linear, polynomial, or periodic (seasonal) mean, or stochastic as for integrated processes of given order, for instance a random walk, which is integrated of order 1, see Section 8.5.

In the multivariate case, deterministic trends can be treated similarly to the univariate case. Therefore we will only consider processes with zero mean here for simplicity. The goal of this chapter is to investigate some particular kind of *stochastic trends* in multivariate time series that have been very popular in multivariate financial time series, giving raise to the concept of *cointegrated* time series.

9.1 VAR processes with integrated variables

Consider the VAR(p) equation, for $\mathbb{C}^m$-valued process $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$,

$$\mathbf{X}_t = \sum_{k=1}^{p} \Phi_k \mathbf{X}_{t-k} + \mathbf{Z}_t \ ,$$

where $\mathbf{Z}$ is a weakly stationary process, for instance a white noise. Using the polynomial

$$\mathbf{\Phi}(z) = \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k \ ,$$

we can compactly write the VAR(p) equation as

$$\mathbf{\Phi}(\mathrm{B})\mathbf{X} = \mathbf{Z} \ .$$

In Section 6.8, we mainly considered the case where $\mathbf{\Phi}$ is stable, in the sense that it remains non singular on the unit disk $\{z \in \mathbb{C} \ : |z| \leq 1\}$. In Sections 8.5 and 8.7 we considered the

case where $\mathbf{\Phi}(1)$ can be singular but only in the case $m = 1$, so that singular simply means zero. If this happens, we basically explained that one can get rid of this case by differencing the time series as many times as the multiplicity order of the root 1, calling this order the *order of integration* of X. It is important to establish this order precisely.

The case $m \geq 2$ is much more involved. Clearly the order of integration of each component can be different. Also one can imagine that specific linear combinations of the vector process behave very differently.

Let us first determine a way to find an order d such that each component $\mathbf{X}(\ell)$ of the process is of integrated order of at most d.

For any square matrix A let us denote by $\hat{A}$ the matrix of its cofactors, that is

$$\tilde{A}(i,j) = (-1)^{i+j} \det\left(A(-i,-j)\right) \ ,$$

where $A(-i,-j)$ is the matrix obtained from A without the i^{th} row and the j^{th} column. Then we have

$$\tilde{A}^T A = \det(A)\ \mathbf{1} \ .$$

In particular, we get that, for $z \in \mathbb{C}$,

$$\tilde{\mathbf{\Phi}}^T(z)\, \mathbf{\Phi}(z) = \det\left(\mathbf{\Phi}(z)\right)\ \mathbf{1} \ .$$

Moreover, we note that $\tilde{\mathbf{\Phi}}^T(z)$ is a (matrix valued) polynomial of z with degree at most $p(m-1)$ such that $\tilde{\mathbf{\Phi}}^T(0) = \mathbf{1}$ and it is non-singular whenever $\mathbf{\Phi}(z)$ is non-singular. Similarly $\det\left(\mathbf{\Phi}(z)\right)$ is a (complex valued) polynomial with degree at most pm, which takes value 1 at 0 and which is non-zero whenever $\mathbf{\Phi}(z)$ is non-singular. Hence we get

$$\det\left(\mathbf{\Phi}(B)\right)\mathbf{X} = \tilde{\mathbf{\Phi}}^T(B)\mathbf{Z} \ ,$$

which is in fact a VARMA equation of the form

$$\mathbf{X}_t(\ell) = \sum_{k=1}^{pm} \phi_k \mathbf{X}_{t-k}(\ell) + \mathbf{Z}_t(\ell) + \sum_{k=1}^{p(m-1)} \left[\Theta_k \mathbf{Z}_{t-k}\right](\ell), \qquad \ell = 1, \ldots, m \ .$$

Note in particular that the AR part of this equation is constant over all $\ell = 1, \ldots, m$. We thus obtain the following result.

Theorem 9.1.1. *Let $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ be a process satisfying the VAR(p) equation*

$$\mathbf{X}_t = \sum_{k=1}^{p} \Phi_k \mathbf{X}_{t-k} + \mathbf{Z}_t \ .$$

Suppose that

$$\mathbf{\Phi}(z) = \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k \ ,$$

is non-singular for all $z \in \mathbb{C}$ such that $|z| \leq 1$ and $z \neq 1$. Let $d \in \mathbb{N}$ be the order of multiplicity of the root 1 in the polynomial $z \mapsto \det(\mathbf{\Phi}(z))$. Then there exists $\tilde{\phi}_1, \ldots, \tilde{\phi}_{pm-d}$ such that

$$1 - \sum_{k=1}^{pm-d} \tilde{\phi}_k z^k \neq 0 \quad \textit{for all} \quad |z| \leq 1$$

and matrices $\tilde{\Theta}_1, \dots, \tilde{\Theta}_{p(m-1)+d}$ *such that, for all* $\ell = 1, \dots, m$,

$$(\mathbf{1} - \mathrm{B})^d \mathbf{X}_t(\ell) = \sum_{k=1}^{pm-d} \tilde{\phi}_k (\mathbf{1} - \mathrm{B})^d \mathbf{X}_{t-k}(\ell) + \mathbf{Z}_t(\ell) + \sum_{k=1}^{p(m-1)+d} \left[\tilde{\Theta}_k \mathbf{Z}_{t-k}\right](\ell) \ .$$

In particular, when $\mathbf{X}$ *has weakly stationary increments and* $\mathbf{Z}$ *is weakly stationary, all components of* $\mathbf{X}$ *are integrated processes of order at most* d.

As previously said, the order d obtained in this theorem may be too large to provide the correct integration order. It may even be larger than the maximal integration order over the components, as illustrated in the following elementary example.

Example 9.1.1. *Let* $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ *be a 2-dimensional process defined such that its two components are two uncorelated ARIMA(0,1,0) processes. Then the order* d *given by Theorem 9.1.1 is* $d = 2$. *(See Exercise 9.1).*

9.2 Definition of cointegrated processes

Consider a bi-dimensional process $\mathbf{X}$ with both components being integrated processes of order 1. It might in fact happen that a linear combination of $\mathbf{X}$ is not integrated. An example is detailed in Exercise 6.6. This means that the integration order 1 of both components can be explained by writing theses components as two different linear combinations of the same two processes, one which is integrated of order one and one which is stationary. We say that the two components of $\mathbf{X}$ are *cointagrated* of order (1,1), abbreviated as $\mathbf{X}$ is CI(1,1). The concept of *cointegration* were introduced to describe this phenomenon in Granger [1981] through several examples. Here we will use the following definition.

Definition 9.2.1 (Cointagration)**.** *Let* $\mathbf{X}$ *be* $\mathbb{C}^m$*-valued process. If all components of* $\mathbf{X}$ *are integrated of order* d *and there exists* $\mathbf{x} \in \mathbb{C}^m$ *such that* $\langle \mathbf{X}, \mathbf{x} \rangle$ *is integrated of order* $d - b$*, we say that* $\mathbf{X}$ *is* cointegrated of order (d, b) *or is CI*(d, b)*. The vector* $\mathbf{x}$ *is called a* cointegrating vector*.*

Let us introduce a standard example.

Example 9.2.1 (Commodity prices)**.** *Cointegrated time series are frequently used to model the prices of a commodity in different markets. Take for instance the case of two markets and let* $\mathbf{X}(1)$ *and* $\mathbf{X}(2)$ *denote the times series of the prices of a given commodity on these two markets. On equilibrium the two prices are linearly related through* $\mathbf{X}(1) = \beta \mathbf{X}(2)$*. Then the increments of the prices are modeled through a linear relation with respect to the equilibrium error*

$$Y = \mathbf{X}(1) - \beta \mathbf{X}(2)$$

Namely,

$$\begin{aligned}(\mathbf{1} - \mathrm{B})\mathbf{X}(1) &= -\alpha_1 \, \mathrm{B}\, Y + \mathbf{Z}(1) \\ (\mathbf{1} - \mathrm{B})\mathbf{X}(2) &= \alpha_2 \, \mathrm{B}\, Y + \mathbf{Z}(2) \ ,\end{aligned}$$

where $\mathbf{Z}$ *is a centered white noise with covariance matrix* $\sigma^2 \mathbf{1}$ *and* α_1, α_2 *and* β *are non-negative coefficients. Setting*

$$\boldsymbol{\alpha} = \begin{bmatrix} -\alpha_1 \\ \alpha_2 \end{bmatrix} \quad \textit{and} \quad \boldsymbol{\beta} = \begin{bmatrix} 1 \\ -\beta \end{bmatrix} \ ,$$

we easily get the VAR(1) equation

$$\mathbf{X}_t = \left(\mathbf{1} + \boldsymbol{\alpha}\boldsymbol{\beta}^T\right) \mathbf{X}_{t-1} + \mathbf{Z}_t \ ,$$

while $Y = \boldsymbol{\beta}^T\mathbf{X}$ *satisfies the AR(1) equation*

$$Y_t = (1 + \boldsymbol{\beta}^T\boldsymbol{\alpha})Y_{t-1} + \boldsymbol{\beta}^T\mathbf{Z}_t \ .$$

If Y *is weakly stationary (the condition* $|1 - \alpha_1 - \beta\alpha_2| < 1$ *holds) and both components of* $\mathbf{X}$ *are I(1) processes. Hence* $\mathbf{X}$ *is CI(1,1).*

9.3 Vector error correlation models (VECM)

In the following, we denote the differencing operator by

$$\Delta = \mathbf{1} - \mathrm{B} \ .$$

Take the process $\mathbf{X}$ of Example 9.2.1. Then its VAR(1) equation can be rewritten as

$$\Delta\,\mathbf{X} = \boldsymbol{\alpha}\boldsymbol{\beta}^T\,\mathrm{B}\,\mathbf{X} + \mathbf{Z} \ . \tag{9.1}$$

This equation is at first sight surprising because in this example the components of $\mathbf{X}$ are I(1) processes with minimal trend degree, so $\Delta\,\mathbf{X}$ and $\mathbf{Z}$ are weakly stationary while $\mathbf{X}$ is not, although they appear in the same linear equation. The apparent contradiction vanishes when realizing that

$$\boldsymbol{\beta}^T\,\mathrm{B}\,\mathbf{X} = \mathrm{B}\,\boldsymbol{\beta}^T\mathbf{X} = \mathrm{B}\,Y$$

and we precisely have that Y is weakly stationary. Note also that for Relation (9.1) to make sense for the CI(1,1) process $\mathbf{X}$ because $\boldsymbol{\alpha}\boldsymbol{\beta}^T$ cannot be replaced by an invertible matrix, otherwise, we could write $\mathrm{B}\,\mathbf{X}(1)$ as a linear combination of weakly stationary processes, which would contradict the fact that it is of integration order 1.

Writing $\mathbf{X}$ as a solution of (9.1) makes it what we call a *vector error correlation models* (VECM) or order 1. This definition extends to the order $p \geq 2$ as follows.

Definition 9.3.1 (Vector error correlation models (VECM))**.** *Let* $p \in \mathbb{N}^*$*. A* $\mathbb{C}^m$*-valued process* $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ *is said to be* vector error correlation model *(VECM) of order* p *if it satisfies the equation*

$$\Delta\,\mathbf{X} = \Pi\,\mathrm{B}\,\mathbf{X} + \sum_{k=2}^{p} \tilde{\Phi}_k\,\mathrm{B}^{k-1}\,\Delta\,\mathbf{X} + \mathbf{Z} \ , \tag{9.2}$$

where Π *and* $\tilde{\Phi}_k$*,* $k = 2, \ldots, p$*, are* $m \times m$ *matrices and* $\mathbf{Z}$ *is a centered white noise.*

In fact, any VAR(p) process can be written as a VECM of order p, since it is just a rewriting of an equation of the form

$$\Phi(\mathrm{B})\mathbf{X} = \mathbf{Z} \Longleftrightarrow \mathbf{X} = \sum_{k=1}^{p} \Phi_k\,\mathrm{B}^k\,\mathbf{X} + \mathbf{Z} \ .$$

Indeed, using that, for $k = 2, \ldots, p$,

$$\mathrm{B}^k = -\,\mathrm{B}^{k-1}\,\Delta + \mathrm{B}^{k-1}$$

the VAR(p) equation is equivalent to (9.2) with

$$\tilde{\Phi}_k = -\left(\Phi_k + \cdots + \Phi_p\right) , \qquad 2 \leq k \leq p , \tag{9.3}$$

$$\Pi = \Phi_1 + \cdots + \Phi_p - \mathbf{1} . \tag{9.4}$$

Equivalently and more concisely said, the above equivalence corresponds to the matrix polynomial equation

$$\mathbf{\Phi}(z) = \mathbf{1} - (\mathbf{1} + \Pi)z - \sum_{k=2}^{p} \tilde{\Phi}_k z^{k-1}(1-z) , \qquad z \in \mathbb{C} . \tag{9.5}$$

The reason why VECM are used in the context of cointegration is that they are adapted to the case where $\mathbf{X}$ have I(0) or I(1) components. First observe that Π is related to the existence of a unit root in Φ, since

$$\mathbf{\Phi}(1) = -\Pi .$$

Suppose that there is a unit root. The extreme such case is when $\Pi = \mathbf{0}$ (the unit root 1 has maximal order in $\det(\Phi(z))$), then

$$\mathbf{\Phi}(z) = (1-z)\tilde{\Phi}(z)$$

with

$$\tilde{\mathbf{\Phi}}(z) = \mathbf{1} - \sum_{k=2}^{p} \tilde{\Phi}_k z^{k-1} ,$$

and $\Delta\,\mathbf{X}$ is solution of the VAR($p-1$) equation

$$\tilde{\mathbf{\Phi}}(\mathrm{B})(\Delta\,\mathbf{X}) = \mathbf{Z} .$$

The case we are interested in is between the two extreme cases : Π invertible, no unit roots *versus* $\Pi = \mathbf{0}$, $\Delta\,\mathbf{X}$ is a VAR($p-1$) process.

Definition 9.3.2 (Cointegration rank)**.** *The m-dimensional weakly stationary increments VECM $\mathbf{X}$ satisfying (9.2) is said to be* cointegrated of rank r *if Π is or rank r. It follows that we can write (see Exercise 9.2)*

$$\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}^H ,$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $m \times r$ full rank matrices called the loading matrix *and the* cointegration matrix*, respectively. They can be taken as real matrices when Π is a real matrix.*

9.4 Filtering non-weakly stationary time series

In the theory of cointegration, we deal with non-stationary solutions of VAR or VECM equations. In the econometric literature, such time series are only defined on positive time indices $\mathbb{N}$ with some initial condition but the initial conditions are not looked at very closely.

Here, in order to get rid of initial conditions, we prefer to see such times series as defined over $\mathbb{Z}$, but with $\mathbf{X}$ and $\mathbf{Z}$ possibly arbitrary on $\mathbb{Z}_-$, provided that they satisfy the following.

Definition 9.4.1 (L^2-backwardly bounded processes)**.** *We will say that a $\mathbb{C}^d$-valued process $\mathbf{X} = (bX_t)_{t\in\mathbb{Z}}$ is L^2-backwardly bounded if it is an L^2 process and*

$$\sup_{t\in\mathbb{N}} \mathbb{E}\left[\left|\mathbf{X}_{-t}\right|^2\right] < \infty .$$

Obviously, if $\mathbf{X}$ is L^2-backwardly bounded, then, for all $t \in \mathbb{Z}$,

$$\sup_{s \le t} \mathbb{E}\left[|\mathbf{X}_s|^2\right] < \infty \ .$$

Adapting Propositions 6.3.3 and 6.3.4, we easily obtain the following result.

Proposition 9.4.1. *Let d, e, f be positive integers. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a L^2-backwardly bounded $\mathbb{C}^d$-valued process. Let $A = (A_n)_{n\in\mathbb{N}}$ and $B = (B_n)_{n\in\mathbb{N}}$ be causal sequences of $e \times d$ matrices and $f \times e$ matrices, respectively, satisfying the absolute summability condition*

$$\sum_{t\in\mathbb{N}} \|A_t\| < \infty \ , \ \sum_{t\in\mathbb{N}} \|B_t\| < \infty \ . \tag{9.6}$$

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

$$Y_t = \sum_{s\in\mathbb{N}} A_s X_{t-s} \ , \tag{9.7}$$

is absolutely convergent in $\mathbb{C}^e$ a.s. and in $L^2(\Omega, \mathbb{C}^e, \mathcal{F}, \mathbb{P})$). Moreover, the process $Y = (Y_t)_{t\in\mathbb{Z}}$ is L^2-backwardly bounded and we have

$$\mathrm{F}_B(Y) = \mathrm{F}_B \circ \mathrm{F}_A(X) = \mathrm{F}_{B\star A}(X) \ ,$$

where $B \star A$ is the sequence of $d \times d$ matrices defined by

$$[B \star A]_k = \sum_{j\in\mathbb{N}} B_j \, A_{k-j} = \sum_{j\in\mathbb{N}} B_{k-j} \, A_j \ .$$

9.5 Granger representation theorem

The Granger representation of a VECM with cointegration rank r provides conditions under which such a vector process is, up to a stationary component, a rank $m-r$ linear combination of the m dimensional random walk defined from the innovation process $\mathbf{Z}$.

Our basic assumption is that of a VAR/VECM representation. We summarize it in the following and introduce somùe notation that will be useful.

Assumption 9.5.1. *The following assertions hold.*

(i) Let $p \ge 1$, $\Phi_1, \ldots, \Phi_p$ be $m \times m$ matrices and define Π and $\tilde{\Phi}_2, \ldots, \tilde{\Phi}_p$ by the relation

$$\mathbf{\Phi}(z) = -\Pi z + (1-z)\,\tilde{\mathbf{\Phi}}(z) \ , \qquad z \in \mathbb{C} \ , \tag{9.8}$$

where

$$\mathbf{\Phi}(z) = \mathbf{1} - \sum_{k=1}^{p} \Phi_k z^k \quad \text{and} \quad \tilde{\mathbf{\Phi}}(z) = \mathbf{1} - \sum_{k=2}^{p} \tilde{\Phi}_k z^k \ , \qquad z \in \mathbb{C} \ .$$

(ii) Suppose moreover that, for some $r \in \{0, \ldots, m\}$,

$$\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}^T \ ,$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $m \times r$ full rank matrices. They are assumed to be real if Π is real.

(iii) We denote by $\boldsymbol{\beta}_\perp$ *an* $m \times (m-r)$ *matrix such that*

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta} & \boldsymbol{\beta}_\perp \end{bmatrix}$$

is a unitary matrix, that is, $\tilde{\boldsymbol{\beta}}^H \tilde{\boldsymbol{\beta}} = \mathbf{1}$*, or, equivalently,* $\boldsymbol{\beta}^H \boldsymbol{\beta} = \mathbf{1}$*,* $\boldsymbol{\beta}_\perp^H \boldsymbol{\beta}_\perp = \mathbf{1}$ *and* $\boldsymbol{\beta}^H \boldsymbol{\beta}_\perp = \mathbf{0}$*. Similarly we denote by* $\boldsymbol{\alpha}_\perp$ *an* $m \times (m-r)$ *matrix with column vectors that form an orthonormal basis of* $(\operatorname{Im} \boldsymbol{\alpha})^\perp$*.*

(iv) The process $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ *is a VECM of order* p *satisfying (9.2) for some process* $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}}$ *and they both are* L^2*-backwardly bounded.*

Relation (9.8) is the same as (9.5) and is equivalent to (9.3) and (9.4). Assumption 9.5.1(ii) amounts to say that Π is of rank r, see Exercise 9.2.

Before stating the theorem, we prove some simple lemmas.

Lemma 9.5.1. *Suppose that Assumption 9.5.1(i)(ii)(iii) hold. Then we have, for all* $z \in \mathbb{C}$*,*

$$\boldsymbol{\Phi}(z) = \check{\boldsymbol{\Phi}}(z) \begin{bmatrix} \boldsymbol{\beta}^H \\ (1-z)\boldsymbol{\beta}_\perp^H \end{bmatrix},$$

where $\check{\boldsymbol{\Phi}}(z)$ *is the* $m \times m$ *matrix*

$$\check{\boldsymbol{\Phi}}(z) = \begin{bmatrix} \Big(-\boldsymbol{\alpha} z + (1-z)\tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}\Big) & \tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}_\perp \end{bmatrix}. \tag{9.9}$$

Proof. By Assumption 9.5.1(i)(ii)(iii) and some elementary algebra, we can write

$$\begin{aligned}
\boldsymbol{\Phi}(z) &= -\Pi + (1-z)\tilde{\boldsymbol{\Phi}}(z) \\
&= \Big(-\boldsymbol{\alpha}\boldsymbol{\beta}^H + (1-z)\tilde{\boldsymbol{\Phi}}(z)\Big)\, \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^H \\
&= \Big(-\boldsymbol{\alpha}\boldsymbol{\beta}^H\tilde{\boldsymbol{\beta}} + (1-z)\tilde{\boldsymbol{\Phi}}(z)\tilde{\boldsymbol{\beta}}\Big)\, \tilde{\boldsymbol{\beta}}^H \\
&= \Big(-\boldsymbol{\alpha}\begin{bmatrix}\mathbf{1}_r & \mathbf{0}_{r,m-r}\end{bmatrix} + (1-z)\tilde{\boldsymbol{\Phi}}(z)\tilde{\boldsymbol{\beta}}\Big)\, \tilde{\boldsymbol{\beta}}^H \\
&= \begin{bmatrix}\Big(-\boldsymbol{\alpha} + (1-z)\tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}\Big) & (1-z)\tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}_\perp\end{bmatrix}\, \tilde{\boldsymbol{\beta}}^H \\
&= \Big(-\boldsymbol{\alpha} + (1-z)\tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}\Big)\, \boldsymbol{\beta}^H + (1-z)\tilde{\boldsymbol{\Phi}}(z)\boldsymbol{\beta}_\perp\boldsymbol{\beta}_\perp^H \\
&= \check{\boldsymbol{\Phi}}(z) \begin{bmatrix} \boldsymbol{\beta}^H \\ (1-z)\boldsymbol{\beta}_\perp^H \end{bmatrix}.
\end{aligned}$$

□

An immediate consequence of Lemma 9.5.1 is that

$$\det\boldsymbol{\Phi}(z) = \det\check{\boldsymbol{\Phi}}(z)\, \overline{\det\tilde{\boldsymbol{\beta}}}\, (1-z)^{m-r},$$

and, in particular, since $\left|\det\tilde{\boldsymbol{\beta}}\right| = 1$, the polynomials $\det\boldsymbol{\Phi}$ and $\det\check{\boldsymbol{\Phi}}$ share the same non-unit roots and the order of multiplicity of the root 1 for $\det\boldsymbol{\Phi}$ is that of $\det\check{\boldsymbol{\Phi}}$ added to $m-r$. Hence the assumption

Assumption 9.5.2. *In addition to Assumption 9.5.1, we have that the polynomial* $\det\boldsymbol{\Phi}$ *has no roots in the closed unit disk other than 1 and the order of multiplicity of the unit root is exactly* $m-r$*.*

is equivalent to saying that the polynomial $\det\check{\boldsymbol{\Phi}}$ has no roots in the closed unit disk and we immediately get the following lemma. Note also that if we already know that the polynomial $\det\boldsymbol{\Phi}$ has no roots in the closed unit disk other than 1, then the order of multiplicity of the unit root is exactly $m-r$ if and only if

$$\check{\boldsymbol{\Phi}}(1) = \begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix} \text{ is invertible.}$$

To deal with this specific condition we have the following result.

Lemma 9.5.2. *Suppose that Assumption 9.5.1(i)(ii)(iii) hold. Then* $\begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix}$ *is invertible if and only if*

$$\boldsymbol{\alpha}_\perp^H \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \text{ is invertible.}$$

Moreover we have in this case that

$$\begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix}^{-1} = \begin{bmatrix} -\left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}^H + \left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1}\boldsymbol{\alpha}_\perp^H \\ \left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1}\boldsymbol{\alpha}_\perp^H \end{bmatrix}.$$

Proof. We multiply the matrix $\begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix}$ on the left by

$$\mathbf{1}_m = \begin{bmatrix} \boldsymbol{\alpha}(\boldsymbol{\alpha}^H\boldsymbol{\alpha})^{-1} & \boldsymbol{\alpha}_\perp(\boldsymbol{\alpha}_\perp^H\boldsymbol{\alpha}_\perp)^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}^H \\ \boldsymbol{\alpha}_\perp^H \end{bmatrix}.$$

We get that

$$\begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}(\boldsymbol{\alpha}^H\boldsymbol{\alpha})^{-1} & \boldsymbol{\alpha}_\perp(\boldsymbol{\alpha}_\perp^H\boldsymbol{\alpha}_\perp)^{-1} \end{bmatrix}\begin{bmatrix} -\boldsymbol{\alpha}^H\boldsymbol{\alpha} & \boldsymbol{\alpha}^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \\ \mathbf{0}_{m-r,m-r} & \boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix}$$

is invertible if and only if the second matrix in the product is invertible and thus, if and only if $\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp$ is invertible. Moreover, in this case, this second matrix has inverse

$$\begin{bmatrix} -\left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1} & \left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1} \\ \mathbf{0}_{m-r,m-r} & \left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1} \end{bmatrix}$$

Since the first matrix of the product has inverse $\begin{bmatrix} \boldsymbol{\alpha}^H \\ \boldsymbol{\alpha}_\perp^H \end{bmatrix}$, we get

$$\begin{bmatrix} -\boldsymbol{\alpha} & \tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp \end{bmatrix}^{-1} = \begin{bmatrix} -\left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1} & \left(\boldsymbol{\alpha}^H\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1} \\ \mathbf{0}_{m-r,m-r} & \left(\boldsymbol{\alpha}_\perp^H\tilde{\boldsymbol{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}^H \\ \boldsymbol{\alpha}_\perp^H \end{bmatrix}.$$

The result follows by bloc multiplication. □

We then get the following result using that $z \mapsto \check{\boldsymbol{\Phi}}(z)^{-1}$ is holomorphic on the open ball of maximal radius over which $\det\check{\boldsymbol{\Phi}}$ does not vanish.

Lemma 9.5.3. *Suppose that Assumption 9.5.1(i)(ii)(iii) and Assumption 9.5.2 hold and define the $\mathbb{C} \to \mathbb{C}^{m\times m}$ function $\check{\mathbf{\Phi}}$ by (9.9). Then there exists a causal sequence $\Theta = (\Theta_n)_{n\in\mathbb{N}} \in (\mathbb{C}^{m\times m})^{\mathbb{N}}$ such that*

$$\rho := \limsup_{n\to\infty} \frac{1}{n} \ln \|\Theta_n\| < 1 \; ,$$

and, for all $z \in \mathbb{C}$ with $|z| < \rho^{-1}$,

$$\check{\mathbf{\Phi}}(z)^{-1} = \sum_{k\in\mathbb{N}} \Theta_k z^k \; .$$

Moreover $\Theta_0 = \check{\mathbf{\Phi}}(0)^{-1} = \tilde{\boldsymbol{\beta}}^H$.

Finally we need the following lemma which decompose a causal convolution filter into a multiplication by a matrix added to the composition of another causal filter with the diffencing operator Δ.

Lemma 9.5.4. *Let $\Theta = (\Theta_n)_{n\in\mathbb{N}} \in (\mathbb{C}^{m\times m})^{\mathbb{N}}$ be a causal sequence of matrices such that*

$$\rho := \limsup_{n\to\infty} \frac{1}{n} \ln \|\Theta_n\| < 1 \; .$$

Then for every L^2-backwardly bounded $\mathbb{C}^d$-valued process $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$, we have

$$\mathrm{F}_\Theta(\mathbf{X}) = \mathbf{\Theta}(1)\, \mathbf{X} + \Delta \circ \mathrm{F}_{\tilde{\Theta}}(\mathbf{X})$$

where

$$\mathbf{\Theta}(z) = \sum_{k\in\mathbb{N}} \Theta_k z^k \; ,$$

and $\tilde{\Theta} = \left(\tilde{\Theta}_n\right)_{n\in\mathbb{N}}$ is defined by

$$\tilde{\Theta}_k = \sum_{j=0}^{k} \Theta_j - \mathbf{\Theta}(1) \; , \qquad k \in \mathbb{N} \; ,$$

or, equivalently, for all $z \in \mathbb{C}$ such that $|z| < \rho^{-1}$,

$$\mathbf{\Theta}(z) = \mathbf{\Theta}(1) + \tilde{\mathbf{\Theta}}(z)(1-z) \; .$$

We can now state the Granger representation theorem.

Theorem 9.5.5 (Granger representation theorem)**.** *Suppose that Assumption 9.5.1(i)(ii)(iii) and Assumption 9.5.2 hold. Then there exists a causal sequence $\check{\Theta} = \left(\check{\Theta}\right)_{n\in\mathbb{N}}$ such that*

$$\rho := \limsup_{n\to\infty} \frac{1}{n} \ln \|\check{\Theta}_n\| < 1 \; ,$$

and the following equation holds.

$$\Delta\, \mathbf{X} = \boldsymbol{\beta}_\perp \left(\boldsymbol{\alpha}_\perp^H \tilde{\mathbf{\Phi}}(1) \boldsymbol{\beta}_\perp\right)^{-1} \boldsymbol{\alpha}_\perp^H\, \mathbf{Z} + \Delta\, \mathrm{F}_{\check{\Theta}}(\mathbf{Z}) \; , \tag{9.10}$$

where the inverse is well defined.

Proof. By Assumption 9.5.1 (iv), and rewriting (9.2) using operators, we have

$$\left(-\Pi\,\mathrm{B}+\tilde{\mathbf{\Phi}}(\mathrm{B})(\mathbf{1}-\mathrm{B})\right)\mathbf{X}=\mathbf{Z}\,.$$

With (9.8) and using Lemma 9.5.1, we obtain

$$\check{\mathbf{\Phi}}(\mathrm{B})\begin{bmatrix}\boldsymbol{\beta}^H\\(1-\mathrm{B})\boldsymbol{\beta}_\perp^H\end{bmatrix}\mathbf{X}=\mathbf{Z}\,.$$

Applying Lemma 9.5.3 and proposition 9.4.1, then gives that

$$\begin{bmatrix}\boldsymbol{\beta}^H\\(1-\mathrm{B})\boldsymbol{\beta}_\perp^H\end{bmatrix}\mathbf{X}=\mathrm{F}_\Theta(\mathbf{Z})\,.$$

This can be separated in two equations, namely,

$$\boldsymbol{\beta}^H\mathbf{X}=\begin{bmatrix}\mathbf{1}_r & \mathbf{0}_{r,m-r}\end{bmatrix}\mathrm{F}_\Theta(\mathbf{Z}) \tag{9.11}$$

$$\boldsymbol{\beta}_\perp^H\,\Delta\,\mathbf{X}=\begin{bmatrix}\mathbf{0}_{m-r,r} & \mathbf{1}_{m-r}\end{bmatrix}\mathrm{F}_\Theta(\mathbf{Z})\,. \tag{9.12}$$

Now writing $\Delta\,\mathbf{X}=\boldsymbol{\beta}\boldsymbol{\beta}^H\,\Delta\,\mathbf{X}+\boldsymbol{\beta}_\perp\boldsymbol{\beta}_\perp^H\,\Delta\,\mathbf{X}$ and using (9.12), we obtain

$$\Delta\,\mathbf{X}=\boldsymbol{\beta}\boldsymbol{\beta}^H\,\Delta\,\mathbf{X}+\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\mathrm{F}_\Theta(\mathbf{Z})\,.$$

We deduce that we have

$$\Delta\,\mathbf{X}=\Delta\,\boldsymbol{\beta}\boldsymbol{\beta}^H\mathbf{X}+\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\mathrm{F}_\Theta(\mathbf{Z})\,.$$

Inserting (9.12), we thus finally get that

$$\Delta\,\mathbf{X}=\begin{bmatrix}\boldsymbol{\beta} & \mathbf{0}_{m,m-r}\end{bmatrix}\Delta\,\mathrm{F}_\Theta(\mathbf{Z})+\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\mathrm{F}_\Theta(\mathbf{Z})\,.$$

It remains to write F_Θ in a more convenient way by using Lemma 9.5.4, which gives that, with the same definition of the causal sequence $\tilde{\Theta}$,

$$\Delta\,\mathbf{X}=\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\mathrm{F}_\Theta(1)\,\mathbf{Z}+\begin{bmatrix}\boldsymbol{\beta} & \mathbf{0}_{m,m-r}\end{bmatrix}\Delta\,\mathrm{F}_\Theta(\mathbf{Z})+\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\Delta\,\mathrm{F}_{\tilde{\Theta}}(\mathbf{Z})\,.$$

The value of $\mathrm{F}_\Theta(1)$ is given in Lemma 9.5.3 as that of $\check{\mathbf{\Phi}}(1)^{-1}$, yielding, with 9.5.2,

$$\begin{aligned}\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\mathrm{F}_\Theta(1)&=\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\check{\mathbf{\Phi}}(1)^{-1}\\&=\begin{bmatrix}\mathbf{0}_{m,r} & \boldsymbol{\beta}_\perp\end{bmatrix}\begin{bmatrix}-\boldsymbol{\alpha} & \tilde{\mathbf{\Phi}}(1)\boldsymbol{\beta}_\perp\end{bmatrix}^{-1}\\&=\boldsymbol{\beta}_\perp\left(\boldsymbol{\alpha}_\perp^H\tilde{\mathbf{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1}\boldsymbol{\alpha}_\perp^H\,.\end{aligned}$$

Equation (9.10) easily follows. □

By Exercise 9.3, the matrix

$$\Xi=\boldsymbol{\beta}_\perp\left(\boldsymbol{\alpha}_\perp^H\tilde{\mathbf{\Phi}}(1)\boldsymbol{\beta}_\perp\right)^{-1}\boldsymbol{\alpha}_\perp^H$$

has rank exactly $m-r$. Hence the Granger representation theorem shows that $\mathbf{X}$ can be seen, up to the additive term $\mathrm{F}_{\check{\Theta}}(\mathbf{Z})$, which is a stable (causal) convolution filter of $\mathbf{Z}$, and up to an additive constant (often called the *initial conditions*), as the rank $m-r$ matrix Ξ applied to the integrated process $I_0(\mathbf{Z})$.

9.6 Exercises

Exercise 9.1. Define $\mathbf{X}(0)$ and $\mathbf{X}(1)$ as two processes satisfying AR(1) equation with AR coefficients $\phi_1, \phi_2 \in (-1, 1]$ and uncorrelated white noise processes.

1. To which case does the process of Example 9.1.1 correspond ?
2. Express the VAR(1) equation satisfied by $\mathbf{X}$

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{Z}_t \ ,$$

   where $\mathbf{Z}$ is a white noise and Φ is a 2×2 matrix to be determined. Check that the assumptions Theorem 9.1.1 are satisfied.
3. Compute $\det(\mathbf{\Phi}(z))$ as defined in Theorem 9.1.1 and determine the order d defined in this theorem.
4. Compare d with the integration orders of $\mathbf{X}(0)$ and $\mathbf{X}(1)$.

Exercise 9.2. Let $m \geq 1$ and A be an $m \times m$ matrix of rank $0 < k \leq m$. The goal of this exercise is to determine all pairs of $m \times k$ matrices α and β such that

$$A = \alpha \, \beta^H \ . \tag{9.13}$$

We first look for necessary conditions on such α and β.

1. Relate $\operatorname{Im} \alpha$ and $\operatorname{Im} \beta$ to $\operatorname{Im} A$ and $\ker A$ and deduce that they have full rank.

We now restrict our search for α and β by adding the condition

$$\beta^H \beta = \mathbf{1} \ . \tag{9.14}$$

Let $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ be two orthonormal sequences in $\mathbb{C}^m$ such that

$$(\ker A)^\perp = \operatorname{Span}(u_1, \ldots, u_k) \quad \text{and} \quad \operatorname{Im} A = \operatorname{Span}(v_1, \ldots, v_k) \ .$$

Set

$$\beta_0 = \begin{bmatrix} u_1 & \ldots & u_k \end{bmatrix} \quad \text{and} \quad \alpha_0 = \begin{bmatrix} v_1 & \ldots & v_k \end{bmatrix} \ .$$

2. Show that $\beta_0 \beta_0^H = \beta \beta^H$.
3. Exhibit a solution of the form $\beta = \beta_0$ and $\alpha = \alpha_0 \, P_0$ for some $k \times k$ matrix P_0.
4. Deduce that the set of all possible pairs (α, β) such that (9.13) and (9.14) hold is given by

$$\{(\alpha_0 \, P_0 \, U, \beta_0 \, U) \ : \ U \ k \times k \text{ unitary matrix}\} \ .$$

5. Show that if A is real, then α and β can be taken real.
6. Deduce that the set of all possible pairs (α, β) such that (9.13) holds is given by

$$\left\{(\alpha_0 \, P_0 \, (P^H)^{-1}, \beta_0 \, P) \ : \ P \ k \times k \text{ invertible matrix } \right\} \ .$$

Exercise 9.3. Under the assumptions of Theorem 9.5.5, determine the kernel and the image of

$$\boldsymbol{\beta}_\perp \left(\boldsymbol{\alpha}_\perp^H \tilde{\boldsymbol{\Phi}}(1) \boldsymbol{\beta}_\perp \right)^{-1} \boldsymbol{\alpha}_\perp^H$$

and deduce that its rank is $m - r$.

Exercise 9.4. In Theorem 9.5.5, if Assumption 9.5.1(iii) is replaced by the following:

(iii′) We denote by $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$ two $m \times (m-r)$ matrices with column vectors that form an orthonormal basis of $(\operatorname{Im} \boldsymbol{\alpha})^\perp$ and $(\operatorname{Im} \boldsymbol{\beta})^\perp$, respectively.

(that is, $\tilde{\boldsymbol{\beta}}$ is not necessarily taken unitary), how should we adapt the conclusion of the theorem ?

Part III

Appendices

Appendix A

Hilbert spaces

Basic knowledge of Hilbert spaces is quite useful for time series, and, more generally stochastic modeling. Here we gather some essential definitions and results on Hilbert spaces. Most results are elementary. A detailed account on this topic can be found in Young [1988].

A.1 Definitions

Definition A.1.1 (Inner-product spaces)**.** *Let $\mathcal{H}$ be a complex linear space. An* Inner-product *on $\mathcal{H}$ is a function*

$$\langle \cdot, \cdot \rangle : x, y \in \mathcal{H} \times \mathcal{H} \mapsto \langle x, y \rangle \in \mathbb{C}$$

which satisfies the following properties

(i) for all $(x, y) \in \mathcal{H} \times \mathcal{H}$, $\langle x, y \rangle = \overline{\langle y, x \rangle}$

(ii) for all $(x, y) \in \mathcal{H} \times \mathcal{H}$ and all $(\alpha, \beta) \in \mathbb{C} \times \mathbb{C}$, $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

(iii) for all $x \in \mathcal{H}$, $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$.

Then the application

$$\| \cdot \| : x \in \mathcal{H} \mapsto \sqrt{\langle x, x \rangle} \geq 0$$

defines a norm on $\mathcal{H}$.

Example A.1.1 ($\mathbb{C}^n$)**.** *The space of column vectors $x = [x_1 \quad \cdots \quad x_n]^T$, where $x_k \in \mathbb{C}$ is a linear space on which the application*

$$\langle x, y \rangle = y^H x = \sum_{k=1}^{n} x_k \overline{y_k}$$

defines an inner product.

Example A.1.2 (ℓ^2)**.** *The space of complex-valued sequences $\{x_k\}_{k \in \mathbb{N}}$ such that $\sum_{k=0}^{\infty} |x_k|^2 < \infty$ is a linear space. Define for all x et y in this space,*

$$\langle x, y \rangle = \sum_{k=0}^{\infty} x_k \overline{y_k} \, .$$

The sum is well defined and finite since $|x_k \overline{y_k}| \leq (|x_k|^2 + |y_k|^2)/2$. Properties (i-iii) of definition A.1.1 are easily verified. We thus obtained an inner-product space, denoted as ℓ^2.

Example A.1.3 (Squared integrable functions)**.** *The space $\mathcal{L}^2(T)$ of $\mathbb{C}$-valued Borel functions defined on an interval $T \subset \mathbb{R}$ whose modulus is squared integrable ($\int_T |f(t)|^2 dt < \infty$) is a linear space. Define*

$$(f,g) \in \mathcal{L}^2(T) \times \mathcal{L}^2(T) \mapsto \langle f,g \rangle = \int_T f(t)\overline{g(t)}\mathrm{d}t \ .$$

As for ℓ^2, Properties (i) and (ii) of Definition A.1.1 hold. However Property (iii) fails to hold since :

$$\langle f,f \rangle = 0 \not\Rightarrow \forall t \in T \ f(t) = 0$$

Instead it implies $f = 0$ a.-e. (almost-everywhere). As a consequence, the space $\mathcal{L}^2(T)$ endowed with $\langle \cdot,\cdot \rangle$ is not an inner-product space. Nevertheless the space $L^2(T)$ of the equivalence classes of $\mathcal{L}^2(T)$ for the a.e. equality is an inner-product space.

Example A.1.4 (Finite variance random variables)**.** *As in Example A.1.3, for all probability space $(\Omega, \mathcal{F}, \mathbb{P})$, one defines $\mathcal{H} = \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ (denoted by $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ if no possible confusion occurs) as the space of all complex-valued random variables X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

$$\mathbb{E}\left[|X|^2\right] < \infty \ .$$

Define moreover

$$(X,Y) \in \mathcal{L}^2(\Omega) \times \mathcal{L}^2(\Omega) \mapsto \langle X,Y \rangle = \mathbb{E}\left[X\overline{Y}\right] \ .$$

For the same reasons as in Example A.1.3, we define the inner-product space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ (or simply L^2 if no ambiguity occurs) as the space of the equivalence classes of $\mathcal{L}^2(\Omega)$ for the a.s. equality. This example and Example A.1.3 can be extended to all measured space $(\Omega, \mathcal{F}, \mu)$ by setting

$$(f,g) \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu) \times \mathcal{L}^2(\Omega, \mathcal{F}, \mu) \mapsto \langle f,g \rangle = \int f\, \overline{g}\ \mathrm{d}\mu \ .$$

We have the following result.

Theorem A.1.1. *For all $x, y \in \mathcal{H} \times \mathcal{H}$, we have :*

a*) Cauchy-Schwarz Inequality:* $|\langle x,y \rangle| \leq \|x\|\|y\|$,

b*) Triangular inequality:* $|\|x\| - \|y\|| \leq \|x-y\| \leq \|x\| + \|y\|$,

c*) Parallelogram inequality:*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

Definition A.1.2 (Convergence in $\mathcal{H}$)**.** *Let (x_n) be a sequence included in an inner-product space $\mathcal{H}$ and $x \in \mathcal{H}$. We say that (x_n) converges to x in $\mathcal{H}$ if $\|x_n - x\| \to 0$ as $n \to +\infty$. We will denote $x_n \to x$ if no confusion occurs with another convergence.*

It is easy to show that, for all $y \in \mathcal{H}$, the application $\langle \cdot, y \rangle : \mathcal{H} \to \mathbb{C}$, $x \mapsto \langle x,y \rangle$ is a continuous linear form. In fact we have the following continuity result.

Theorem A.1.2 (Continuity of the inner product)**.** *If $x_n \to x$ and $y_n \to y$ in the inner-product space $\mathcal{H}$, then $\langle x_n, y_n \rangle \to \langle x,y \rangle$. In particular, $\|x_n\| \to \|x\|$.*

Proof. Using the triangle inequality and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\langle x, y\rangle - \langle x_n, y_n\rangle &= \langle (x - x_n) + x_n, (y - y_n) + y_n\rangle - \langle x_n, y_n\rangle \\ &= \langle x - x_n, y - y_n\rangle + \langle x - x_n, y_n\rangle + \langle x_n, y - y_n\rangle \\ &\leq \|x_n - x\|\|y_n - y\| + \|x_n - x\|\|y_n\| + \|y_n - x\|\|x_n\|\end{aligned}$$

This implies the result, since the sequences (x_n) are (y_n) bounded. □

Definition A.1.3 (Hilbert space)**.** *An inner-product space $\mathcal{H}$ is called an* Hilbert space *if it is complete (that is, every Cauchy sequence converges).*

Recall that a normed space is complete if and only if every absolutely convergent series is convergent see [Royden, 1988, Proposition 5 in Chapter 6, Page 124].

Example A.1.5 (ℓ^2)**.** *The space ℓ^2 is a Hilbert space. Let (a_n) be a Cauchy sequence in ℓ^2. Denote*

$$a_n = (a_{n,1}, a_{n,2}, \dots) ,$$

then, for all $\epsilon > 0$, there exists N such that, for all $n, m \geq N$,

$$\sum_{k=1}^{\infty} |a_{m,k} - a_{n,k}| \leq \epsilon^2 . \tag{A.1}$$

Let k be fixed. The previous displays shows that $(a_{n,k})_n$ is a Cauchy sequence in $\mathbb{C}$. Let α_k denote its limit. Further denote $a = (\alpha_k)$. It remains to show that $a \in \ell^2$ and $\lim_{n\to\infty} \|a_n - a\| = 0$. Using (A.1)*, we have for all $p \in \mathbb{N}$, and all $m, n \geq N$,*

$$\sum_{k=1}^{p} |a_{m,k} - a_{n,k}|^2 \leq \sum_{k=1}^{\infty} |a_{m,k} - a_{n,k}|^2 \leq \epsilon^2.$$

Hence, for all $p \in \mathbb{N}$ and all $n \geq N$, $\lim_{m\to\infty} \sum_{k=1}^{p} |a_{m,k} - a_{n,k}|^2 = \sum_{k=1}^{p} |\alpha_k - a_{n,k}|^2 \leq \epsilon^2$. Taking the limit as $p \to \infty$, we thus get, for all $n \geq N$,

$$\|a - a_n\|^2 = \sum_{k=1}^{\infty} |\alpha_k - a_{n,k}|^2 \leq \epsilon^2,$$

which implies $(a - a_n) \in \ell^2$, thus $a \in \ell^2$. Since ϵ is arbitrary, we also get that $\lim_{n\to\infty} \|a - a_n\| = 0$.

Proposition A.1.3 (L^2 spaces)**.** *For all measured space $(\Omega, \mathcal{F}, \mu)$, the space $L^2(\Omega, \mathcal{F}, \mu)$ (see Example A.1.4) endowed with*

$$\langle f, g\rangle = \int f\bar{g} \, \mathrm{d}\mu$$

is a Hilbert space.

A more general result on L^p spaces is given in [Royden, 1988, Proposition 6 in Chapter 6, Page 126].

Example A.1.6 (A non-complete inner-product space)**.** *Let $\mathcal{C}([-\pi,\pi])$ the space of continuous functions on $[-\pi,\pi]$. It is a subspace of the Hilbert space $L^2([-\pi,\pi])$. However it is not closed since $\mathbb{1}_{[-\pi/2,\pi/2]}$ can be approximated by continuous functions with arbitrarily small L^2 error. Hence $\mathcal{C}([-\pi,\pi])$ endowed with*

$$\langle f,g\rangle = \int_{-\pi}^{\pi} f\bar{g}$$

is not a complete space, although it is an inner product space.

Definition A.1.4 (Generated subspace and its closure)**.** *Let $\mathcal{X}$ be a subspace of $\mathcal{H}$. We denote by* Span $(\mathcal{X})$ *the subspace of all finite linear combinations of vectors in $\mathcal{X}$ and by* $\overline{\text{Span}}\,(\mathcal{X})$ *the* closure *of* Span $(\mathcal{X})$ *in $\mathcal{H}$, that is the smallest closed subspace of $\mathcal{H}$ that contains* Span $(\mathcal{X})$*. In fact* $\overline{\text{Span}}\,(\mathcal{X})$ *contains and only contains all elements of $\mathcal{H}$ which are L^2 limits of sequences included in* Span $(\mathcal{X})$*.*

Definition A.1.5 (Orthogonality)**.** *Two vectors $x,y\in\mathcal{H}$ are* orthogonal*, if $\langle x,y\rangle = 0$, which we denoted by $x\perp y$. If $\mathcal{S}$ is a subspace of $\mathcal{H}$, we write $x\perp\mathcal{S}$ if $x\perp s$ for all $s\in\mathcal{S}$. Also we write $\mathcal{S}\perp\mathcal{T}$ if all vectors in $\mathcal{S}$ are orthogonal to $\mathcal{T}$.*

Take two subspaces $\mathcal{A}$ et $\mathcal{B}$ such that $\mathcal{H}=\mathcal{A}+\mathcal{B}$, that is, for all $h\in\mathcal{H}$, there exist $a\in\mathcal{A}$ et $b\in\mathcal{B}$ such that $h=a+b$. If moreover $\mathcal{A}\perp\mathcal{B}$ we will denote $\mathcal{H}=\mathcal{A}\overset{\perp}{\oplus}\mathcal{B}$.

Definition A.1.6 (Orthogonal set)**.** *Let $\mathcal{E}$ be a subset of an Hilbert space $\mathcal{H}$. The orthogonal set of $\mathcal{E}$ is defined as*

$$\mathcal{E}^{\perp} = \{x\in\mathcal{H} : \forall y\in\mathcal{E}\ \ \langle x,y\rangle = 0\}$$

We will need the following result, whose proofs are left to the reader as an exercise.

Theorem A.1.4. *If $\mathcal{E}$ is a subset of an Hilbert space $\mathcal{H}$, then $\mathcal{E}^{\perp}$ is closed.*

A.2 Orthogonal and orthonormal bases

Definition A.2.1 (Orthogonal and orthonormal sets)**.** *Let E be a subset of $\mathcal{H}$. It is an orthogonal set if for all $(x,y)\in E\times E$, $x\neq y$, $\langle x,y\rangle = 0$. If moreover $\|x\|=1$ for all $x\in E$, we say that E is orthonormal.*

Linear combinations of vectors in an orthogonal set have the following remarkable property. Let E be an orthogonal set and $x_1,\dots,x_n\in E$ distinct. Then for all $(\alpha_1,\dots,\alpha_n)\in\mathbb{C}^n$,

$$\left\|\sum_{k=1}^{n}\alpha_k x_k\right\|^2 = \sum_{k=1}^{n}|\alpha_k|^2\|x_k\|^2\,. \tag{A.2}$$

Thus the vectors of an orthogonal set are linearly independent. Relation (A.2) is well known in Euclidean geometry. In Hilbert spaces, we can extend this formula to infinite sums.

Theorem A.2.1. *Let $(e_i)_{i\geq 1}$ be an orthonormal sequence of an Hilbert space $\mathcal{H}$ and let $(\alpha_i)_{i\geq 1}$ be a sequence of complex numbers. The series*

$$\sum_{i=1}^{\infty}\alpha_i e_i \tag{A.3}$$

converges in $\mathcal{H}$ if and only if $\sum_i |\alpha_i|^2 < \infty$, in which case

$$\left\| \sum_{i=1}^{\infty} \alpha_i e_i \right\|^2 = \sum_{i=1}^{\infty} |\alpha_i|^2 . \tag{A.4}$$

Proof. For all $m > k > 0$, as in (A.2), we have

$$\left\| \sum_{i=k}^{m} \alpha_i e_i \right\|^2 = \sum_{i=k}^{m} |\alpha_i|^2 .$$

since $\sum_{i=1}^{\infty} |\alpha_i|^2 < \infty$, the sequence $s_m = \sum_{i=1}^{m} \alpha_i e_i$ is Cauchy in $\mathcal{H}$. Since $\mathcal{H}$ is complete, it converges. Relation (A.4) is obtained by taking the limit.

Conversely, if $\sum_{i=1}^{\infty} \alpha_i e_i$ is convergent series, then (A.4) again holds, which implies the converse result. □

An Orthonormal series allows us to approximate any $x \in \mathcal{H}$ by a finite partial sum of the infinite sum (A.3).

Proposition A.2.2. *Let x be a vector of the Hilbert space $\mathcal{H}$ and $E = \{e_1, \cdots, e_n\}$ a finite orthonormal set of vectors, then*

$$\left\| x - \sum_{k=1}^{n} \langle x, e_k \rangle e_k \right\|^2 = \|x\|^2 - \sum_{k=1}^{n} |\langle x, e_k \rangle|^2 . \tag{A.5}$$

In addition, $\sum_{k=1}^{n} \langle x, e_k \rangle e_k$ is the vector of $\mathrm{Span}(e_1, \ldots, e_n)$ *that is the closest of x. Hence the left-hand side of (A.5) equals*

$$\inf \left\{ \|x - y\|^2 \ : \ y \in \mathrm{Span}(e_1, \ldots, e_n) \right\} .$$

Proof. We have for all $j = 1, \ldots, n$,

$$\left\langle x - \sum_{k=1}^{n} \langle x, e_k \rangle e_k, e_j \right\rangle = \langle x, e_j \rangle - \langle x, e_j \rangle = 0 .$$

Hence, we may write

$$x = \left(x - \sum_{k=1}^{n} \langle x, e_k \rangle e_k \right) + \sum_{i=1}^{n} \langle x, e_k \rangle e_k ,$$

which is the sum of two orthogonal vectors. By Pythagore's Identity and (A.2) with $x_k = e_k$ and $\alpha_k = \langle x, e_k \rangle$, we get (A.5).

Similarly,, for all $(\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$,

$$\left\| x - \sum_{k=1}^{n} \alpha_k e_k \right\|^2 = \left\| x - \sum_{k=1}^{n} \langle x, e_k \rangle e_k \right\|^2 + \sum_{k=1}^{n} |\langle x, e_k \rangle - \alpha_k|^2 ,$$

and thus $\sum_{k=1}^{n} \langle x, e_k \rangle e_k$ achieves the best approximation of x by a linear combinations of $e_1, \ldots, e_n$. □

Example A.2.1 (Gram-Schmidt algorithm)**.** *Let $(y_i)_{i\geq 1}$ be a sequence in a Hilbert space $\mathcal{H}$. The Gram-Schmidt algorithm is an iterative algorithm to construct an orthogonal sequence such that* $\mathrm{Span}\,(e_1,\ldots,e_n) = \mathrm{Span}\,(y_1,\ldots,y_n)$ *for all $n \geq 1$.*

Algorithm 12: Gram-Schmidt algorithm.

Data: A set of vectors $y_1,\ldots,y_n$
Result: An orthogonal sequence $e_1,\ldots,e_n$
Initialization: set $e_1 = y_1$.
for $t = 2,\ldots,n$ **do**
  Define
$$e_t = y_t - \sum_{k=1}^{t-1} \frac{\langle y_t, e_k\rangle}{\|e_k\|^2} e_k \ ,$$
  with the convention $0/0 = 0$.
end

Proposition A.2.2 also yields the following result.

Corollary A.2.3 (Bessel Inequality)**.** *Let $(e_i)_{i\geq 1}$ be an orthonormal sequence of a Hilbert space $\mathcal{H}$. Then, for all $x \in \mathcal{H}$,*

$$\sum_{i=1}^{\infty} |\,\langle x, e_i\rangle\,|^2 \leq \|x\|^2 \ .$$

The Bessel inequality implies that for all $x \in \mathcal{H}$, $\lim_{n\to\infty} \langle x, e_n\rangle = 0$ and also that $(\langle x, e_i\rangle)_{i\geq 1}$ is in ℓ^2. By Theorem A.2.1, we get that

$$\sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i \tag{A.6}$$

is a convergent series. It is called the *Fourier expansion* of x; the coefficients $\langle x, e_i\rangle$ are called the *Fourier coefficients* with respect to the orthonormal sequence (e_i). Note however that, although $\sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i$ always converges, its limit is not always equal to x.

Example A.2.2. *Let $\mathbb{T}$ denote the quotient space $\mathbb{R}/(2\pi\mathbb{Z})$ (or any interval congruent to $[0, 2\pi)$). Consider $\mathcal{H} = L^2(\mathbb{T})$ and define $e_n(t) = \pi^{-1/2}\sin(nt)$ pour $n = 1, 2, \cdots$. The sequence (e_n) is orthonormal in $\mathcal{H}$, but for $x(t) = \cos(t)$, we have*

$$\sum_{n=1}^{\infty} \langle x, e_n\rangle\, e_n(t) = \sum_{n=1}^{\infty} \left[\pi^{-1/2} \int_{\mathbb{T}} \cos(t)\sin(nt)\mathrm{d}t\right] \pi^{-1/2}\sin(nt)$$
$$= \sum_{n=}^{\infty} 0 \cdot \sin(nt) = 0 \neq \cos t \ .$$

In fact the limit is x, if an additional property is assumed.

Definition A.2.2 (Dense sets, Hilbert Bases)**.** *A subset E of a Hilbert space $\mathcal{H}$ is said* dense *if $\overline{\mathrm{Span}}\,(E) = \mathcal{H}$. An* orthonormal dense *sequence is called a* Hilbert basis.

Let us give an example of a dense set for measured spaces.

Proposition A.2.4. *Consider the measured space $(\Omega, \mathcal{F}, \mu)$ and the Hilbert space $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mu)$.*

$$\overline{\text{Span}}\,(\mathbb{1}_A, A \in \mathcal{F}) = L^2(\Omega, \mathcal{F}, \mu)\ ,$$

Proof. For any nonnegative square integrable function f defined on $(\Omega, \mathcal{F}, \mu)$, denote

$$f_n = \sum_{k=0}^{n2^n} k2^{-n} \mathbb{1}_{f^{-1}([k2^{-n},(k+1)2^{-n}))} \in \overline{\text{Span}}\,(\mathbb{1}_A, A \in \mathcal{F})\ .$$

Since $0 \le f_n \le f$, by dominated convergence, we get that $\int |f_n - f|^2 \mathrm{d}\mu \to 0$. Since any $g \in L^2(\Omega, \mathcal{F}, \mu)$ is a linear combination of at most 4 nonnegative functions (the positive and negative part of the real and complex parts), we get the result. □

A Hilbert basis allows us to "reach" any point in $\mathcal{H}$.

Theorem A.2.5. *Let $(e_i)_{i\ge1}$ be a Hilbert basis of the Hilbert space $\mathcal{H}$. Then for all $x \in \mathcal{H}$,*

$$x = \sum_{i=1}^{\infty} \langle x, e_i\rangle\, e_i\ . \tag{A.7}$$

Proof. We already known the series in (A.6) converges. On the other hand, since (e_i) is dense, there exists $(\alpha_{p,n})_{1\le i\le n}$ such that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \alpha_{i,n} e_i = x\ .$$

Now from Proposition A.2.2, we have

$$\left\| x - \sum_{i=1}^{n} \langle x, e_i\rangle\, e_i \right\| \le \left\| x - \sum_{i=1}^{n} \alpha_{i,n} e_i \right\|\ .$$

Hence the result. □

Theorem A.2.5 implies that an orthonormal sequence (e_i) is a Hilbert basis if and only if Relation (A.7) holds for all $x \in \mathcal{H}$. The proof of the following result is left as an exercise.

Theorem A.2.6. *Let $(e_i)_{i\ge1}$ be an orthonormal sequence of the Hilbert space $\mathcal{H}$. The following assertions are equivalent.*

(i) $(e_i)_{i\ge1}$ *is an Hilbert basis.*

(ii) If some $x \in \mathcal{H}$ satisfies

$$\langle x, e_i\rangle = 0 \quad \textit{for all} \quad i \ge 1\ ,$$

then $x = 0$.

(iii) For all $x \in \mathcal{H}$,

$$\|x\|^2 = \sum_{i=1}^{\infty} |\langle x, e_i\rangle|^2\ . \tag{A.8}$$

Example A.2.3 (Fourier basis)**.** *Define*

$$e_n(x) = (2\pi)^{-1/2} \mathrm{e}^{\mathrm{i}nx} \, , n \in \mathbb{Z} \, .$$

Then (e_n) *is an Hilbert basis of* $L^2(\mathbb{T})$*, see e.g. Young [1988].*

A Hilbert space is called separable if it contains a countable dense subset.

Theorem A.2.7. *A Hilbert space* $\mathcal{H}$ *is separable if and only if it admits a Hilbert basis.*

Proof. Let (e_i) be a Hilbert basis of $\mathcal{H}$. The set $S = \bigcup_{n=1}^{\infty} S_n$, where, for $n \in \mathbb{N}$,

$$S_n \stackrel{\text{def}}{=} \left\{ \sum_{l=1}^{n} (\alpha_k + \mathrm{i}\beta_k) e_k \, , (\alpha_k, \beta_k) \in \mathbb{Q} \times \mathbb{Q}, k = 1, \ldots, n \right\}$$

is countable. Since for $x \in \mathcal{H}$,

$$\lim_{n \to \infty} \left\| \sum_{k=1}^{n} \langle x, e_k \rangle \, e_k - x \right\| = 0 \, ,$$

the set S is dense in $\mathcal{H}$.

If $\mathcal{H}$ is separable then there exists a dense sequence $(y_i)_{i \geq 1}$. The Gram-Schmidt algorithm of Example A.2.1 provides an orthogonal sequence $(e_i)_{i \geq 1}$ such that $\mathrm{Span}\,(e_1, \ldots, e_n) = \mathrm{Span}\,(y_1, \ldots, y_n)$ for all n. Removing the null vectors in this sequence and normalizing the others by the square root of their norms, we get a Hilbert basis. □

A.3 Fourier series

Define the sequence of complex exponential functions

$$\phi_n(x) = (2\pi)^{-1/2} \mathrm{e}^{\mathrm{i}nx}, \quad n \in \mathbb{Z} \, . \tag{A.9}$$

We shall see that (ϕ_n) is a dense set of $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ for any finite measure μ on the Borel sets of $\mathbb{T}$. If moreover μ est the Lebesgue mesure, it is a Hilbert basis.

Let $L^1(\mathbb{T})$ denote the set of 2π-periodic locally integrable (with respect to the Lebesgue measure) functions. For $f \in L^1(\mathbb{T})$, set

$$f_n = \sum_{k=-n}^{n} \left(\int_{\mathbb{T}} f \bar{\phi}_k \right) \phi_k \, , \quad n = 0, 1, 2, \ldots$$

Then

$$f_n(x) = \sum_{k=-n}^{n} \frac{1}{2\pi} \int_{\mathbb{T}} f(t) \mathrm{e}^{\mathrm{i}k(x-t)} \mathrm{d}t \, . \tag{A.10}$$

The following result can be found in Young [1988].

Theorem A.3.1. *Suppose that* f *is a continuous* 2π*-periodic function. Then the Cesaro sequence*

$$\left(\frac{1}{n} \sum_{k=0}^{n-1} f_k \right)_{n \in \mathbb{N}^*}$$

converges uniformly to f*.*

An interesting consequence for us is the following result (see Exercise A.2).

Corollary A.3.2. *Let μ be a finite measure on the Borel sets of $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$. The sequence $(\phi_n)_{n\in\mathbb{Z}}$ defined in (A.9) is linearly dense in the Hilbert space $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$, that is, $\overline{\mathrm{Span}}\,(\phi_n,\, n \in \mathbb{Z}) = L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$.*

In the case of the Lebesgue measure, we get the following.

Corollary A.3.3. *The sequence $(\phi_n)_{n\in\mathbb{Z}}$ defined in (A.9) is a Hilbert basis in $L^2(\mathbb{T})$. In particular, for all $f \in L^2(\mathbb{T})$,*

$$f = \sum_{k=-\infty}^{\infty} \alpha_k \phi_k \quad \textit{with} \quad \alpha_k = (2\pi)^{-1/2} \int_{\mathbb{T}} f(x) \mathrm{e}^{-\mathrm{i}kx} \mathrm{d}x \,,$$

where the infinite sum converges in $L^2(\mathbb{T})$. The Parseval identity then reads

$$\int_{\mathbb{T}} |f(x)|^2 \mathrm{d}x = \sum_{k=-\infty}^{\infty} |\alpha_k|^2 \,.$$

A.4 Projection and orthogonality principle

The following theorem allows us to define the orthogonal projection onto a closed subspace of a Hilbert space.

Theorem A.4.1 (Projection theorem)**.** *Let $\mathcal{E}$ be a closed convex subset of a Hilbert space $\mathcal{H}$ and let $x \in \mathcal{H}$. Then the following assertions hold.*

(i) There exists a unique vector $\mathrm{proj}\,(x|\,\mathcal{E}) \in \mathcal{E}$ *such that*

$$\|x - \mathrm{proj}\,(x|\,\mathcal{E})\,\| = \inf_{w\in\mathcal{E}} \|x - w\|$$

(ii) If moreover $\mathcal{E}$ is a linear subspace, $\mathrm{proj}\,(x|\,\mathcal{E})$ *is the unique $\hat{x} \in \mathcal{E}$ such that $x - \hat{x} \in \mathcal{E}^\perp$.*

We call $\mathrm{proj}\,(x|\,\mathcal{E})$ *the orthogonal projection of x onto $\mathcal{E}$.*

Proof. Let $x \in \mathcal{H}$. Set $h = \inf_{w\in\mathcal{E}} \|x - w\| \geq 0$. Let $w_1, w_2, \cdots$, be in $\mathcal{E}$ such that:

$$\lim_{m\to+\infty} \|x - w_m\|^2 = h^2 \geq 0 \tag{A.11}$$

The parallelogram identity $\|a-b\|^2+\|a+b\|^2 = 2\|a\|^2+2\|b\|^2$ with $a = w_m - x$ and $b = w_n - x$ gives that

$$\|w_m - w_n\|^2 + \|w_m + w_n - 2x\|^2 = 2\|w_m - x\|^2 + 2\|w_n - x\|^2$$

Since $(w_m + w_n)/2 \in \mathcal{E}$, we have $\|w_m + w_n - 2x\|^2 = 4\|(w_m + w_n)/2 - x\|^2 \geq 4h^2$. By (A.11), for all $\epsilon > 0$,there exists N such that $\forall m, n > N$:

$$\|w_m - w_n\|^2 \leq 2(h^2 + \epsilon) + 2(h^2 + \epsilon) - 4h^2 = 4\epsilon \,,$$

which shows that $\{w_n, n \in \mathbb{N}\}$ is a Cauchy sequence and thus converges to some limit y in $\mathcal{E}$, since $\mathcal{E}$ is closed. By continuity of the norm, we have $\|y - x\| = h$.

It remains to show the uniqueness. Let $z \in \mathcal{E}$ such that $\|x-z\|^2 = \|x-y\|^2 = h^2$. Then the parallelogram identity implies

$$0 \leq \|y-z\|^2 = -4\|(y+z)/2 - x\|^2 + 2\|x-y\|^2 + 2\|x-z\|^2 \\ \leq -4h^2 + 2h^2 + 2h^2 = 0 \,,$$

where we used that $(y+z)/2 \in \mathcal{E}$ with the convexity assumption and thus $\|(y+z)/2 - x\|^2 \geq h^2$. We get $y = z$, which conclude the proof of this assertion.

We now prove the second assertion. Let $\hat{x}$ be the orthogonal projection of x onto $\mathcal{E}$. If there exists $u \in \mathcal{E}$ such that $x - u \perp \mathcal{E}$, we have

$$\begin{aligned} \|x-\hat{x}\|^2 &= \langle x-u+u-\hat{x}, x-u+u-\hat{x}\rangle \\ &= \|x-u\|^2 + \|u-\hat{x}\|^2 + 2\mathrm{Re}(\langle u-\hat{x}, x-u\rangle) \\ &= \|x-u\|^2 + \|u-\hat{x}\|^2 + 0 \geq \|x-u\|^2 \,, \end{aligned}$$

and thus $u = \hat{x}$ by the previous assertion.

Conversely suppose that $x - \hat{x} \not\perp \mathcal{E}$ and let us find a contradiction. Then there exists $y \in \mathcal{E}$ such that $\|y\| = 1$ and $c = \langle x-\hat{x}, y\rangle \neq 0$. Set $\tilde{x} = \hat{x} + cy \in \mathcal{E}$. We have

$$\begin{aligned} \|x-\tilde{x}\|^2 &= \langle x-\hat{x}+\hat{x}-\tilde{x}, x-\hat{x}+\hat{x}-\tilde{x}\rangle \\ &= \|x-\hat{x}\|^2 + \|\hat{x}-\tilde{x}\|^2 + 2\mathrm{Re}(\langle \hat{x}-\tilde{x}, x-\hat{x}\rangle) \\ &= \|x-\hat{x}\|^2 - |c|^2 < \|x-\hat{x}\|^2 \,. \end{aligned}$$

Thus we get a contradiction with the definition of $\hat{x}$. □

Assertion (ii) provides a quite practical way to determine the projection, since it replaces a minimization problem by a system of linear equations to solve.

Example A.4.1 (Projection onto a one dimension space)**.** *Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{C} = \mathrm{Span}\,(v)$ with $v \in \mathcal{H}$. For any $x \in \mathcal{H}$, we have $\mathrm{proj}\,(x|\,\mathcal{C}) = \alpha v$ with $\alpha = \langle x, v\rangle / \|v\|^2$. Denoting $\epsilon = x - \mathrm{proj}\,(x|\,\mathcal{C})$, we get*

$$\|\epsilon\|^2 = \|x\|^2\left(1 - \|\rho\|^2\right) \quad \textit{where} \quad \rho = \frac{\langle x, v\rangle}{\|x\|\|v\|} \quad \textit{with} \quad |\rho| \leq 1$$

The projection operator defined by Theorem A.4.1 has the following interesting properties, whose proofs are left to the reader as an exercise.

Proposition A.4.2. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{E}$ a closed subspace of $\mathcal{H}$. Then the following assertions hold.*

(i) Suppose that $\mathcal{E} = \overline{\mathrm{Span}}\,(e_k, k \in \mathbb{N})$ with (e_k) being an orthonormal sequence. Then

$$\mathrm{proj}\,(h|\,\mathcal{E}) = \sum_{k=0}^{\infty} \langle h, e_k\rangle\, e_k \,.$$

(ii) The function $\mathrm{proj}\,(\cdot|\,\mathcal{E}) : \mathcal{H} \to \mathcal{H}$, $x \mapsto \mathrm{proj}\,(x|\,\mathcal{E})$ is linear and continuous on $\mathcal{H}$.

(iii) $\|x\|^2 = \|\,\mathrm{proj}\,(x|\,\mathcal{E})\,\|^2 + \|x - \mathrm{proj}\,(x|\,\mathcal{E})\,\|^2$,

(iv) $x \in \mathcal{E}$ *if and only if* $\mathrm{proj}\,(x|\,\mathcal{E}) = x$.

(v) $x \in \mathcal{E}^{\perp}$ *if and only if* $\mathrm{proj}\,(x|\,\mathcal{E}) = 0$.

(vi) Let $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two closed subspace of* $\mathcal{H}$*, such that* $\mathcal{E}_1 \subset \mathcal{E}_2$*. Then*

$$\forall x \in \mathcal{H}, \qquad \mathrm{proj}\,(\mathrm{proj}\,(x|\,\mathcal{E}_2)|\,\mathcal{E}_1) = \mathrm{proj}\,(x|\,\mathcal{E}_1)\ .$$

(vii) Let $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two closed subspace of* $\mathcal{H}$*, such that* $\mathcal{E}_1 \perp \mathcal{E}_2$*. Then*

$$\forall x \in \mathcal{H}, \qquad \mathrm{proj}\left(x|\,\mathcal{E}_1 \overset{\perp}{\oplus} \mathcal{E}_2\right) = \mathrm{proj}\,(x|\,\mathcal{E}_1) + \mathrm{proj}\,(x|\,\mathcal{E}_2)\ .$$

The following result will be useful.

Theorem A.4.3. *Let* $(\mathcal{M}_n)_{n\in\mathbb{Z}}$ *be an increasing sequence of closed subspaces of an Hilbert space* $\mathcal{H}$*.*

(i) Denote $\mathcal{M}_{-\infty} = \bigcap_n \mathcal{M}_n$*. Then for all* $h \in \mathcal{H}$*, we have*

$$\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty}) = \lim_{n\to-\infty} \mathrm{proj}\,(h|\,\mathcal{M}_n)$$

(ii) Let $\mathcal{M}_\infty = \overline{\bigcup_{n\in\mathbb{Z}} \mathcal{M}_n}$*. Then, for all* $h \in \mathcal{H}$*,*

$$\mathrm{proj}\,(h|\,\mathcal{M}_\infty) = \lim_{n\to\infty} \mathrm{proj}\,(h|\,\mathcal{M}_n)\ .$$

Proof. We first note that (ii) can be deduced from (i). Indeed, we have

$$\mathcal{M}_\infty^{\perp} = \bigcap_n \mathcal{M}_n^{\perp}\ ,$$

and thus, since $\mathcal{M}_\infty$ and the $\mathcal{M}_n$'s are closed, Assertion (vii) of Proposition A.4.2 yields $\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty}) = h - \mathrm{proj}\left(h|\,\mathcal{M}_\infty^{\perp}\right)$ and the same holds for $\mathcal{M}_n$. Now, since $\mathcal{M}_n^{\perp}$ are closed by Theorem A.1.4, we can apply (i).

It remains to show (i). Since $\mathcal{M}_n$ is a closed subspace of $\mathcal{H}$, $\mathcal{M}_{-\infty}$ is a closed subspace of $\mathcal{H}$. The projection theorem, Theorem A.4.1, shows that $\mathrm{proj}\,(h|\,\mathcal{M}_{-\infty})$ exists. For $m < n$, define $\mathcal{M}_n \ominus \mathcal{M}_m$ as the orthogonal complement of $\mathcal{M}_m$ in $\mathcal{M}_n$, that is $\mathcal{M}_m^{\perp} \cap \mathcal{M}_n$.This is a closed subset of $\mathcal{H}$ by Theorem A.1.4. Using Assertion (vii) of Proposition A.4.2,

$$\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_m) = \mathrm{proj}\,(h|\,\mathcal{M}_n) - \mathrm{proj}\,(h|\,\mathcal{M}_m)\ .$$

It follows that, for all $m \geq 1$,

$$\sum_{n=-m+1}^{0} \|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2 = \|\,\mathrm{proj}\,(h|\,\mathcal{M}_0 \ominus \mathcal{M}_{-m})\,\|^2 \leq \|h\|^2 < \infty\ .$$

We obtain that the series $(\|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2)_{n\leq 0}$ is convergent and since for all $m \leq p \leq 0$,

$$\|\,\mathrm{proj}\,(h|\,\mathcal{M}_p) - \mathrm{proj}\,(h|\,\mathcal{M}_n)\,\|^2 = \sum_{n=-m+1}^{p} \|\,\mathrm{proj}\,(h|\,\mathcal{M}_n \ominus \mathcal{M}_{n-1})\,\|^2\ ,$$

the sequence $\{\mathrm{proj}\,(h|\,\mathcal{M}_n)\,, n = 0, -1, -2, \dots\}$ is a Cauchy sequence. Since $\mathcal{H}$ is complete, $\mathrm{proj}\,(h|\,\mathcal{M}_n)$ converges in $\mathcal{H}$, say to z. We have to show that $z = \mathrm{proj}\,(h|\,\mathcal{M}_{-\infty})$. By the projection theorem, this is equivalent to $z \in \mathcal{M}_{-\infty}$ and $h - z \perp \mathcal{M}_{-\infty}$. Since $\mathrm{proj}\,(h|\,\mathcal{M}_n) \in \mathcal{M}_p$ for all $n \leq p$, we have $z \in \mathcal{M}_p$ for all p and thus $z \in \mathcal{M}_{-\infty}$. Take now $p \in \mathcal{M}_{-\infty}$. Then $p \in \mathcal{M}_n$ for all $n \in \mathbb{Z}$, and, for all $n \in \mathbb{Z}$, $\langle h - \mathrm{proj}\,(h|\,\mathcal{M}_n)\,, p\rangle = 0$ and $\langle h - z, p\rangle = 0$ by taking the limit, which achieves the proof. □

A.5 Riesz representation theorem

We start with some simple results on the orthogonal set.

Proposition A.5.1. *Let $\mathcal{E}$ and $\mathcal{F}$ be two subspaces of a Hilbert space $\mathcal{H}$. If $\mathcal{E} \overset{\perp}{\oplus} \mathcal{F} = \mathcal{H}$, then $\mathcal{F} = \mathcal{E}^{\perp}$.*

Proof. Any $x \in \mathcal{H}$ can be written as $x = y + z$ with $y \in \mathcal{E}$ and $z \in \mathcal{F} \subseteq \mathcal{E}^{\perp}$. Hence $x \in \mathcal{E}^{\perp}$ if and only if $y \in \mathcal{E}^{\perp}$, and so $y = 0$, and thus $x = z \in \mathcal{F}$. □

However, one needs an additional assumption on $\mathcal{E}$ in order to have that $\mathcal{E} \overset{\perp}{\oplus} \mathcal{F} = \mathcal{H}$ and $\mathcal{F} = \mathcal{E}^{\perp}$ are two equivalent assertions.

Theorem A.5.2. *If $\mathcal{E}$ is a closed subspace of a Hilbert space $\mathcal{H}$, then $\mathcal{E} \overset{\perp}{\oplus} \mathcal{E}^{\perp} = \mathcal{H}$. Moreover $(\mathcal{E}^{\perp})^{\perp} = \mathcal{E}$.*

Proof. Let $x \in \mathcal{H}$ and set $y = \mathrm{proj}\,(x|\,\mathcal{E})$. Then $z = x - y \in \mathcal{E}^{\perp}$ by characterization of the orthogonal projection. Hence $x = y + z$ with $y \in \mathcal{E}$ and $z \in \mathcal{E}^{\perp}$, which shows the first assertion of the theorem.

The second assertion is a consequence of the first one and of Proposition A.5.1. □

We can now state the main result of this section.

Theorem A.5.3 (Riesz representation theorem)**.** *Let $\mathcal{H}$ be a Hilbert space. Then $F : \mathcal{H} \to \mathbb{C}$ is a non-zero continuous linear form if and only if there exists $x \in \mathcal{H} \setminus \{0\}$ such that $F(y) = \langle y, x\rangle$ for all $y \in \mathcal{H}$.*

Proof. Let $x \in \mathcal{H} \setminus \{0\}$ and $F : \mathcal{H} \to \mathbb{C}$ defined by $F(y) = \langle y, x\rangle$ for all $y \in \mathcal{H}$. Then F is a continuous linear form by linearity of the scalar product with respect to the first argument and by the Cauchy-Schwarz inequality. Moreover F is non-zero since $F(x) > 0$.

Let us now show the direct implication. Let $F : \mathcal{H} \to \mathbb{C}$ be a non-zero continuous linear form. Denote by $\mathcal{E}$ the null space of F. Then $\mathcal{E}$ is a closed subspace of $\mathcal{H}$. By Theorem A.5.2, $\mathcal{E}^{\perp}$ is a supplementary set of $\mathcal{E}$ in $\mathcal{H}$. Since $\mathcal{E}$ has codimension 1, we conclude that $\mathcal{E}^{\perp}$ has dimension 1. Let $z \in \mathcal{E}^{\perp}$ such that $\|z\| = 1$, hence $\mathcal{E}^{\perp} = \mathrm{Span}\,(z)$. Then for all $y \in \mathcal{H}$, we have $\mathrm{proj}\,(y|\,\mathrm{Span}\,(z)) = \langle y, z\rangle\, z$ and, since $y - \mathrm{proj}\,(y|\,\mathrm{Span}\,(z)) \in \mathcal{E} = (\mathcal{E}^{\perp})^{\perp}$, we get $F(y) = \langle y, z\rangle\, F(z)$. We conclude the proof by setting $x = \overline{F(z)}z$. □

A.6 Unitary operators

Definition A.6.1 (Unitary operators)**.** *Let $\mathcal{H}$ and $\mathcal{I}$ be two Hilbert spaces. An* isometric operator *S from $\mathcal{H}$ to $\mathcal{I}$ is a linear application $S : \mathcal{H} \to \mathcal{I}$ such that $\langle Sv, Sw\rangle_{\mathcal{I}} = \langle v, w\rangle_{\mathcal{H}}$ for*

all $(v, w) \in \mathcal{H}$. *If it is moreover bijective, we say that it is a* unitary operator. *In this case we also say that* $\mathcal{H}$ *and* $\mathcal{I}$ *are isomorphic.*

Observe that an isometric operator is always continuous.

Theorem A.6.1. *Let* $\mathcal{H}$ *be a separable Hilbert space.*

(i) If $\mathcal{H}$ *has infinite dimension, it is isomorphic to* ℓ^2.

(ii) If $\mathcal{H}$ *has dimension* n, *it is isomorphic to* $\mathbb{C}^n$.

Proof. It is a direct application of Theorem A.2.7 and Theorem A.2.1. □

The following result is very convenient to construct isometric operators.

Theorem A.6.2. *Let* $(\mathcal{H}, \langle , \rangle_{\mathcal{H}})$ *and* $(\mathcal{I}, \langle , \rangle_{\mathcal{I}})$ *be two Hilbert spaces. Let* $\mathcal{G}$ *be a subspace of* $\mathcal{H}$.

(i) Let $S : \mathcal{G} \to \mathcal{I}$ *be isometric on* $\mathcal{G}$. *Then* S *admits a unique isometric extension* $\bar{S} : \overline{\mathcal{G}} \to \mathcal{I}$ *and* $\bar{S}(\overline{\mathcal{G}})$ *is the closure of* $S(\mathcal{G})$ *in* $\mathcal{I}$.

(ii) Let $(v_t, t \in \mathrm{T})$ *and* $(w_t, t \in \mathrm{T})$ *be two sets of vectors in* $\mathcal{H}$ *and* $\mathcal{I}$ *indexed by an arbitrary index set* T. *Suppose that for all* $(s, t) \in \mathrm{T} \times \mathrm{T}$, $\langle v_t, v_s \rangle_{\mathcal{H}} = \langle w_t, w_s \rangle_{\mathcal{I}}$. *Then, there exists a unique isometric operator* $S : \overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T}) \to \overline{\mathrm{Span}}\,(w_t, t \in \mathrm{T})$ *such that for all* $t \in \mathrm{T}$, $Sv_t = w_t$. *Moreover,* $S\left(\overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T})\right) = \overline{\mathrm{Span}}\,(w_t, t \in \mathrm{T})$.

One often uses the same notation for S and its extension $\bar{S}$.

Proof. We first show Assertion (i). Let $v \in \bar{\mathcal{G}}$. For all sequence $(v_n) \subset \mathcal{G}$ converging to v, the sequence (Sv_n) is a Cauchy sequence in $\mathcal{I}$ (since (v_n) is Cauchy in $\mathcal{G}$ and S is isometric). Thus there exists $w \in \mathcal{I}$ such that $w = \lim_{n\to\infty} Sv_n$. If (v'_n) is also converging to v, we have $\|v'_n - v'_n\|_{\mathcal{H}} \to 0$ and thus $\|Sv_n - Sv'_n\|_{\mathcal{I}} \to 0$, which shows that w only depends on v. Set $\bar{S}v = w$. Linearity and isometric properties are preserved by taking the limit and $\bar{S} : \bar{\mathcal{G}} \to \mathcal{I}$ is thus an isometric extension of S. The uniqueness of this extension is obvious.

By definition $\bar{S}(\bar{\mathcal{G}})$ is included in the closure of $S(\mathcal{G})$. Conversely, let $w \in \overline{S(\mathcal{G})}$. there exists a sequence $(v_n) \in \mathcal{G}$ such that $w = \lim_{n\to\infty} Sv_n$. The sequence (Sv_n) is Cauchy and thus so is (v_n) in $\mathcal{G}$. Let $v \in \bar{\mathcal{G}}$ its limit. We have $\bar{S}v = \lim_{n\to\infty} Sv_n$ and thus $\bar{S}v = w$, which shows that $\overline{S(\mathcal{G})} \subseteq \bar{S}(\bar{\mathcal{G}})$. The first assertion is proved.

We next show the second assertion. For all finite subset J of T and all complex numbers $(a_t)_{t\in J}$ and $(b_t)_{t\in J}$, we have

$$\sum_{t\in J} a_t v_t = \sum_{t\in J} b_t v_t \Rightarrow \sum_{t\in J} a_t w_t = \sum_{t\in J} b_t w_t$$

since by setting $c_t = a_t - b_t$,

$$\left\| \sum_{t\in J} c_t v_t \right\|^2_{\mathcal{H}} = \sum_{t\in J} \sum_{t'\in J} c_t \overline{c_{t'}} \left\langle v_t, v'_t \right\rangle_{\mathcal{H}} = \sum_{t\in J} \sum_{t'\in J} c_t \overline{c_{t'}} \left\langle w_t, w'_t \right\rangle_{\mathcal{I}} = \left\| \sum_{t\in I} c_t w_t \right\|_{\mathcal{I}},$$

using the linearity and isometric properties. This allows us to define $Sf = \sum_{t\in I} a_t w_t$ for all f such that $f = \sum_{t\in I} a_t v_t$ with I finite subset of T. We just defined S on $\mathcal{G} = \mathrm{Span}\,(v_t, t \in \mathrm{T})$ and it is an isometric operator. Applying (i), it admits a unique isometric extension $\bar{S} : \bar{\mathcal{G}} \to \mathcal{I}$ such that $\bar{S}(\bar{\mathcal{G}}) = \overline{S(\mathcal{G})}$. By definition, $\bar{\mathcal{G}} = \overline{\mathrm{Span}}\,(v_t, t \in \mathrm{T})$ and $S(\mathcal{G}) = \mathrm{Span}\,(w_t, t \in \mathrm{T})$. □

A.7 Exercises

Exercise A.1. Let X and Y be two complex valued random variables in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, for some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Determine the constant $m = \mathrm{proj}\,(X | \,\mathrm{Span}\,(1))$.
2. Determine the random variable $Z = \mathrm{proj}\,(X | \,\mathrm{Span}\,(1, Y))$.

Exercise A.2 (The Fourier basis is dense). The first questions of this exercise are dedicated to the proof of Theorem A.3.1. Let f be a continuous 2π-periodic function and f_n be defined as in (A.10).

1. Determine the Fejér kernel J_n, which satisfies
$$\frac{1}{n}\sum_{k=0}^{n-1} f_k = \int_{\mathbb{T}} J_n(x-t)\, f(t)\, \mathrm{d}t\ .$$
2. Show that we can write, for all $t \in \mathbb{R}$,
$$J_n(t) = \frac{1}{2\pi}\sum_{k=-n+1}^{n-1} (1-|k|/n)\mathrm{e}^{\mathrm{i}kt} = \frac{1}{2\pi n}\left|\sum_{j=0}^{n-1}\mathrm{e}^{\mathrm{i}jt}\right|^2\ .$$
3. Deduce that $J_n \geq 0$, $\int_{\mathbb{T}} J_n = 1$ and that for any $\epsilon \in (0, \pi]$,
$$\sup_{n\geq 1} n \sup_{\epsilon \leq |t| \leq \pi} J_n(t) < \infty\ .$$
4. Conclude the proof of Theorem A.3.1.

Let now μ be a finite measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. Let F be a closed set in $\mathbb{T}$. Define $f_n(x) = (1 - n\, d(F, x))_+$, where $d(F, x) = \inf\{|y - x| \ :\ y \in F\}$.

5. Show that $f_n \to \mathbb{1}_F$ in $L^1(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$.

By Proposition C.1.3, we know that μ is regular that is, for all $A \in \mathcal{B}(\mathbb{T})$,
$$\mu(A) = \inf\{\mu(U) \ :\ U \text{ open set } \supset A\} = \sup\{\mu(F) \ :\ F \text{ closed set} \subset A\}\ .$$

6. Deduce that for all $A \in \mathcal{B}(\mathbb{T})$ and all $\epsilon > 0$, there exists a continuous 2π-periodic function g_ϵ such that
$$\int |\mathbb{1}_A - g_\epsilon| \mathrm{d}\mu \leq \epsilon\ .$$
7. Deduce that the set of continuous 2π-periodic functions is dense in $L^1(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ endowed with the L^1 norm.
8. Deduce that the set of continuous 2π-periodic functions is also dense in $L^2(\mathbb{T}, \mathcal{B}(\mathbb{T}), \mu)$ endowed with the L^2 norm.
9. Conclude the proof of Corollary A.3.2.

Appendix B

Probability

A detailed account on the foundations of probability can be founded in Royden [1988]. Here we first recall some important and useful results for characterizing probability or extending properties of integrals established on specific functions to more general ones. Then two fundamental tools are introduced: conditional expectation and density functions with respect to σ-finite measures. These tools allow us to define conditional distributions and kernels, which will be repeatedly used in the following chapters.

B.1 Useful remainders

We gathered some classical results that we will use repeatedly. The proofs can be omitted as a first reading.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, a non empty space Ω endowed with a σ-field $\mathcal{F}$ and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. Recall that a π-system $\mathcal{C}$ on Ω is a non-empty class of subsets of Ω which is stable by finite intersection, for all $A, B \in \mathcal{C}$, $A \cap B \in \mathcal{C}$. Also recall that a λ-system $\mathcal{A}$ on Ω is a class of subsets of Ω which contains Ω and is stable by taking the complementary set or a countable union of disjoint sets: for all $A \in \mathcal{A}$, $A^c \in \mathcal{A}$, and for all $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ such that $A_n \cap A_p = \emptyset$ for $n < p$, $\cup_n A_n \in \mathcal{A}$.

The following theorem is very useful for extending a result from a π-system to its generating σ-field. Recall that for a class $\mathcal{C}$ of subsets of Ω, we denote by $\sigma(\mathcal{C})$ the smallest σ-field containing $\mathcal{C}$, which is the intersection of all σ-fields containing $\mathcal{C}$.

Theorem B.1.1 ($\pi - \lambda$-theorem)**.** *Let $\mathcal{C} \subset \mathcal{A}$ with $\mathcal{C}$ a π-system and $\mathcal{A}$ a λ-system of the same space Ω. Then $\sigma(\mathcal{C}) \subset \mathcal{A}$.*

The proof of Theorem B.1.1 relies on simple lemmas which we now state.

Lemma B.1.2. *Let $\mathcal{A}$ be a class of subsets of Ω which contains Ω. It is a λ system (as defined above) if and only if the two following assertions are satisfied.*

(i) For all $A, B \in \mathcal{A}$ such that $A \subseteq B$, $B \setminus A \in \mathcal{A}$.

(ii) For all nondecreasing sequence $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$, we have $\cup_n A_n \in \mathcal{A}$.

Proof. See Exercise B.1. □

Lemma B.1.3. *Let $\mathcal{A}$ a be λ-system of the space Ω and let $B \in \mathcal{A}$. Then $\{A \in \mathcal{A} : A \cap B \in \mathcal{A}\}$ is a λ-system of the space Ω.*

Proof. See Exercise B.2. □

Lemma B.1.4. *If $\mathcal{A}$ is a λ-system and a π-system of Ω, then it is a σ-field of Ω.*

Proof. Let $\mathcal{A}$ be a λ-system and a π-system of Ω, Then it is stable by finite intersection, complementary set (and thus also by finite union) and countable union of pairwise disjoint sets. To conclude that $\mathcal{A}$ is a σ-field of Ω, it suffices to take $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ and show that we have $\cup_n A_n \in \mathcal{A}$. Define $B_0 = A_0$ and for all $n \in \mathbb{N}^*$, set

$$B_n := A_n \cap (A_0 \cup \cdots \cup A_{n-1})^c \in \mathcal{A} .$$

Then we have, by an inductive reasoning on $p \in \mathbb{N}$,

$$\bigcup_{n=0}^{p} A_n = \bigcup_{n=0}^{p} B_n .$$

Moreover, the B_n's are pairwise disjoint and thus $\cup_n B_n \in \mathcal{A}$. Thus $\cup_n A_n \in \mathcal{A}$, which concludes the proof. □

Proof of Theorem B.1.1. Let $\mathcal{C} \subset \mathcal{A}$ with $\mathcal{C}$ a π-system and $\mathcal{A}$ a λ-system of the same space Ω. Let $\mathcal{A}'$ be the intersection of all λ-systems containing $\mathcal{C}$, which is a λ-system containing $\mathcal{C}$ (hence the smallest one). It follows that $\mathcal{A}' \subseteq \mathcal{A}$. Hence, to conclude the proof, it suffices to show that $\mathcal{A}' = \sigma(\mathcal{C})$, or, equivalently, that $\mathcal{A}'$ is a σ-field. To this end, let us denote, for all $B \subseteq \Omega$,

$$\mathcal{A}'_B := \left\{ A \in \mathcal{A}' \,:\, A \cap B \in \mathcal{A}' \right\} .$$

By Lemma B.1.3, we have that $\mathcal{A}'_B$ is a λ-system for all $B \in \mathcal{A}'$. Moreover we easily check that, for all $B \in \mathcal{C}$, $\mathcal{C} \subset \mathcal{A}'_B \subset \mathcal{A}'$. Thus we have, by definition of $\mathcal{A}'$, $\mathcal{A}'_B = \mathcal{A}'$ for all $B \in \mathcal{C}$, that is, for all $A \in \mathcal{A}'$ and $B \in \mathcal{C}$, $A \cap B \in \mathcal{A}'$. It follows that, for all $C \in \mathcal{A}'$ and $B \in \mathcal{C}$, we have $B \in \mathcal{A}'_C$. So, for all $C \in \mathcal{A}'$, $\mathcal{A}'_C$ is a λ-system which contains $\mathcal{C}$. Thus, using again the same reasoning, by definition of $\mathcal{A}'$, for all for all $C \in \mathcal{A}'$, $\mathcal{A}'_C = \mathcal{A}'$. This gives us that, for all $C \in \mathcal{A}'$ and $A \in \mathcal{A}'$, $A \cap C \in \mathcal{A}'$. Hence $\mathcal{A}'$ is also a π-system and, as λ-system, by Lemma B.1.4, is is a σ-field, which concludes the proof. □

Characterizing a probability measure from its application over a rich enough subclass of sets is repeatedly used in these lecture notes and it is worthwhile to remind the reader of the way Theorem B.1.1 can be used to make such an argument work.

Theorem B.1.5 (Characterization of probability measures)**.** *Let $\mathcal{C}$ be a π-system on Ω. Then a probability measure μ on $(\Omega, \sigma(\mathcal{C}))$ is uniquely characterized by $\mu(A)$ on $A \in \mathcal{C}$.*

Proof. Take two probability measures μ and μ' on $(\Omega, \sigma(\mathcal{C}))$ such that $\mu(A) = \mu'(A)$ for all $A \in \mathcal{C}$. It is straightforward to check that $\mathcal{A} := \{A \in \sigma(\mathcal{C}) \,:\, \mu(A) = \mu'(A)\}$ is a λ-system. We then conclude by applying Theorem B.1.1 □

Let us also recall a common tool in integration theory, which consists in writing non-negative measurable functions as limits of a non-decreasing sequence of simple functions.

Proposition B.1.6. *Let $(\Omega, \mathcal{F})$ be a measurable space and $f \in \mathrm{F}_+(\Omega, \mathcal{F})$. Define, for all $n \in \mathbb{N}^*$,*

$$f_n = \sum_{k=0}^{n\,2^n-1} k\,2^{-n}\mathbb{1}_{\{k\,2^{-n}\le f<(k+1)2^{-n}\}} + n\,\mathbb{1}_{\{n\le f\}}\,.$$

Then $(f_n(\omega))_{n\in\mathbb{N}^*}$ is a non-decreasing sequence tending to $f(\omega)$ at every point $\omega \in \Omega$.

Proof. Let us set, for all integers $n \ge 1$,

$$\begin{aligned} &A_{k,n} := \left\{k\,2^n \le f < (k+1)2^{-n}\right\} \text{ for all integers } 0 \le k < n\,2^n,\\ &A_{n2^n,n} := \{n \le f\}\,. \end{aligned}$$

Let now $n \in \mathbb{N}^*$. Then, by definition, we have that $(A_{k,n})_{0\le k\le (n+1)\,2^{n+1}}$ is a partition of Ω satisfying

$$\text{for all } k' \in \{0,\dots,n\,2^n-1\},\quad A_{2k',n+1}\cup A_{2k'+1,n+1} = A_{k',n}\,, \tag{B.1}$$

$$\text{and}\quad \left(\bigcup_{k=n\,2^{n+1}}^{(n+1)\,2^{n+1}-1} A_{k,n+1}\right)\cup A_{(n+1)2^{n+1},n+1} = A_{n\,2^n,n}\,. \tag{B.2}$$

We thus get that

$$\begin{aligned} f_{n+1} &= \sum_{k=0}^{(n+1)\,2^{n+1}-1} k\,2^{-n-1}\mathbb{1}_{A_{k,n+1}} + (n+1)\,\mathbb{1}_{A_{(n+1)2^{n+1},n+1}}\\ &= \sum_{k'=0}^{n\,2^n-1}\left((2k')\,2^{-n-1}\mathbb{1}_{A_{2k',n+1}} + (2k'+1)\,2^{-n-1}\mathbb{1}_{A_{2k'+1,n+1}}\right)\\ &\quad + \sum_{k=n\,2^{n+1}}^{(n+1)\,2^{n+1}-1} k\,2^{-n-1}\mathbb{1}_{A_{k,n+1}} + (n+1)\,\mathbb{1}_{A_{(n+1)2^{n+1},n+1}}\\ &\ge \sum_{k'=0}^{n\,2^n-1} k'\,2^{-n}\,\mathbb{1}_{A_{k',n}} + n\mathbb{1}_{A_{n\,2^n,n}} = f_n\,. \end{aligned}$$

Therefore $(f_n(\omega))_{n\in\mathbb{N}^*}$ is a non-decreasing sequence. Moreover, by construction, we have, for all $\omega \in \Omega$, if $f(\omega) < n$, then $f_n(\omega) \le f(\omega) \le f_n(\omega) + 2^{-n}$ and if $f(\omega) \ge n$, then $f_n(\omega) = n$. Separating the cases $f(\omega) < \infty$ and $f(\omega) = \infty$, this yields

$$\lim_{n\to\infty} f_n(\omega) = f(\omega)\,,$$

which concludes the proof. □

We conclude this preamble with a useful consequence of Proposition B.1.6.

Corollary B.1.7. *Let $(\Omega, \mathcal{F})$ and $(\mathsf{X}, \mathcal{X})$ be two measurable spaces and X a measurable function from $(\Omega, \mathcal{F})$ onto $(\mathsf{X}, \mathcal{X})$. Let $\sigma(X) = \left\{X^{-1}(A) : A \in \mathcal{X}\right\}$. Then $\sigma(X)$ is smallest σ-field endowing Ω to make X measurable from Ω onto $(\mathsf{X}, \mathcal{X})$. Let now $f : \Omega \to \mathbb{R}$. Then f is measurable from $(\Omega, \sigma(X))$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ if and only if there exists g measurable from $(\mathsf{X}, \mathcal{X})$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $f = g \circ X$.*

Proof. The fact that $\sigma(X)$ is smallest σ-field endowing Ω to make X measurable from Ω onto $(\mathsf{X}, \mathcal{X})$ is straightforward to check. Also if $f = g \circ X$ with g measurable from $(\mathsf{X}, \mathcal{X})$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then we immediately get that f is measurable from $(\Omega, \sigma(X))$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Let us now assume that f is measurable from $(\Omega, \sigma(X))$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It only remains to show that there exists g measurable from $(\mathsf{X}, \mathcal{X})$ onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $f = g \circ X$. Since we can write $f = f_+ - f_-$ with $f_+ = \max(0, f)$ and $f_- = \max(0, -f)$, we can assume without loss of generality that f is measurable from $(\Omega, \sigma(X))$ onto $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$.

Let $A_{k,n}$ be defined for all $n \in \mathbb{N}^*$ and $0 \leq k \leq n\, 2^n$ as in the proof of Proposition B.1.6. Since f is measurable from $(\Omega, \sigma(X))$ onto $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, for all integers $0 \leq k \leq n\, 2^n$, there exists $B_{k,n} \in \mathcal{X}$ such that

$$A_{k,n} = X^{-1}\left(B_{k,n}\right) \ . \tag{B.3}$$

Now, by definition of f_n in Proposition B.1.6, using (B.3) and the fact that, for all $C \in \mathcal{X}$, $\mathbb{1}_{X^{-1}(C)} = \mathbb{1}_C \circ X$, we get that, for all $n \in \mathbb{N}^*$,

$$f_n = \sum_{k=0}^{n\,2^n - 1} k\, 2^{-n} \mathbb{1}_{A_{k,n}} + n\, \mathbb{1}_{A_{n2^n,n}} = g_n \circ X \ , \tag{B.4}$$

where g_n is the measurable function defined on $(\mathsf{X}, \mathcal{X})$ onto $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ by

$$g_n := \sum_{k=0}^{n\,2^n - 1} k\, 2^{-n} \mathbb{1}_{B_{k,n}} + n\, \mathbb{1}_{B_{n2^n,n}} \ .$$

Now define, for all $n \in \mathbb{N}^*$, $\tilde{g}_n = \max\left(g_1, \ldots, g_n\right)$, so that $\left(\tilde{g}_n\right)_{n \in \mathbb{N}^*}$ is a non-decreasing sequence of functions in $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$. Its limit $\tilde{g}$ therefore belongs to $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$. We further have, using (B.4) and Proposition B.1.6, for all $\omega \in \Omega$,

$$\begin{aligned}
\tilde{g} \circ X(\omega) &= \lim_{n \to \infty} \max\left(g_1\left(X(\omega)\right), \ldots, g_n\left(X(\omega)\right)\right) \\
&= \lim_{n \to \infty} \max\left(f_1(\omega), \ldots, f_n(\omega)\right) \\
&= \lim_{n \to \infty} f_n(\omega) \\
&= f(\omega) \ .
\end{aligned}$$

Defining g as the function equating $\tilde{g}$ on $\{\tilde{g} < \infty\}$ and taking value 0 everywhere else, we thus get that g is measurable from $(\mathsf{X}, \mathcal{X})$ onto $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ and $g \circ X = f$, which concludes the proof. □

B.2 Conditional Expectation

Recall that, for $p \in [1, \infty]$ we denote by $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ the space of random variables X such that $\mathbb{E}\left[|X|^p\right] < \infty$ for $p < \infty$ and $\sup |X| < \infty$ for $p = \infty$, and by $L^p(\Omega, \mathcal{F}, \mathbb{P})$ its quotient with respect to $\mathbb{P}$-a.s. equality.

Recall also that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space and observe that for any sub-σ-field $\mathcal{G}$ of $\mathcal{F}$, the space $L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Thus, by Proposition A.4.2, for any real valued $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, $Y = \mathrm{proj}\left(X \middle| L^2(\Omega, \mathcal{G}, \mathbb{P})\right)$ is well defined and satifies

(i) $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$,

(ii) $\mathbb{E}\left[|X-Y|^2\right] = \inf_{Z \in L^2(\Omega,\mathcal{G},\mathbb{P})} \mathbb{E}\left[|X-Z|^2\right]$.

(iii) For all $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$, $\mathbb{E}\left[(X-Y)Z\right] = 0$.

Moreover (i) and (ii) are sufficient to characterize Y, as well as (i) and (iii). Looking at (iii) we see that it can be written as $\mathbb{E}\left[XZ\right] = \mathbb{E}\left[YZ\right]$ and that this identity continues to make sense if one relaxes the condition $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ into $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and impose $Z \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. In fact, it turns out that taking Z of the form of an indicator function $\mathbb{1}_A$ is sufficient for constructing and defining Y as stated in the following result.

Lemma B.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Then there exists $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ such that*

$$\mathbb{E}\left[X\mathbb{1}_A\right] = \mathbb{E}\left[Y\mathbb{1}_A\right] \quad \text{for all } A \in \mathcal{G}. \tag{B.5}$$

Moreover the following assertions hold.

(i) If $Y' \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ also satisfies (B.5)*, then $Y = Y'$.*

(ii) If $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathrm{proj}\left(X \middle| L^2(\Omega, \mathcal{G}, \mathbb{P})\right) = \left\{Y' \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P}) \,:\, Y' = Y \,\mathbb{P}\text{-a.s.}\right\}$.

Proof. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $X_n = n \wedge (X \vee (-n))$ so that $X_n \to X$ in L^1 with $X_n \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ for all $n \geq 1$. Define $Y_n = \mathrm{proj}\left(X_n \middle| L^2(\Omega, \mathcal{G}, \mathbb{P})\right)$. Then for all $n \geq 1$ and $A \in \mathcal{G}$, we have

$$\mathbb{E}\left[X_n\mathbb{1}_A\right] = \mathbb{E}\left[Y_n\mathbb{1}_A\right] \,. \tag{B.6}$$

Moreover for all $p \geq n \geq 1$, we have $\mathrm{proj}\left(X_n - X_p \middle| L^2(\Omega, \mathcal{G}, \mathbb{P})\right) = Y_n - Y_p$ and so

$$\mathbb{E}\left[|Y_n - Y_p|\right] = \mathbb{E}\left[(Y_n - Y_p)\mathrm{sgn}(Y_n - Y_p)\right] = \mathbb{E}\left[(X_n - X_p)\mathrm{sgn}(Y_n - Y_p)\right] \,,$$

since $\mathrm{sgn}(Y_n - Y_p) \in L^2(\Omega, \mathcal{G}, \mathbb{P})$. Hence we get that

$$\mathbb{E}\left[|Y_n - Y_p|\right] = \left|\mathbb{E}\left[(X_n - X_p)\mathrm{sgn}(Y_n - Y_p)\right]\right| \leq \mathbb{E}\left[|X_n - X_p|\right] \,.$$

We conclude that $(Y_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^1(\Omega, \mathcal{G}, \mathbb{P})$. Its limit Y then satisfies (B.5) by letting $n \to \infty$ in (B.6). Hence we have proven the main assertion of the lemma.

We now prove (i). Let $Z = Y - Y' \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ so that $\mathbb{E}\left[Z\mathbb{1}_A\right] = 0$ for all $A \in \mathcal{G}$. Then taking A successively equal to $\{Z > 0\}$ and $\{Z < 0\}$ we get $Z_+ = Z_- = 0$, thus $Z = 0$.

Finally, to prove (ii), take $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then any $Y' = \mathrm{proj}\left(X \middle| L^2(\Omega, \mathcal{G}, \mathbb{P})\right)$ is in $L^1(\Omega, \mathcal{G}, \mathbb{P})$ and satisfies (B.5). Assertion (ii) then follows from (i). □

Lemma B.2.1 allows us to introduce the following definition.

Definition B.2.1 (Conditional expectation)**.** *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The unique $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ defined by (B.5) is called the* conditional expectation *of X* given *$\mathcal{G}$, and denoted by $Y = \mathbb{E}\left[X \middle| \mathcal{G}\right]$.*

Conditional expectations are defined as elements of L^1, thus they are random variables defined up to $\mathbb{P}$-almost sure equality. Hence, when writing $\mathbb{E}\left[X \middle| \mathcal{G}\right] = Y$ for instance, we always mean that this relations holds for the $\mathbb{P}$-a.s. equality.

We now have a series of simple lemmas.

Lemma B.2.2. *Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The mapping $X \mapsto \mathbb{E}[X|\mathcal{G}]$ is linear continuous from $L^1(\Omega, \mathcal{F}, \mathbb{P})$ to itself. Moreover we have*

$$\mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|] \leq \mathbb{E}[|X|] \ . \tag{B.7}$$

Proof. The linearity of conditional expectation is left as an exercise (see Proposition B.2.5(a) and Exercise B.5). Continuity is then a byproduct of (B.7), which we now prove. Let Y be a $\mathcal{G}$-measurable version of $\mathbb{E}[X|\mathcal{G}]$ and $A = \{Y \geq 0\}$. Then $A, A^c \in \mathcal{G}$ and thus

$$\mathbb{E}[|Y|] = \mathbb{E}[\mathbb{1}_A Y] - \mathbb{E}[\mathbb{1}_{A^c} Y] = \mathbb{E}[\mathbb{1}_A X] - \mathbb{E}[\mathbb{1}_{A^c} X] = \mathbb{E}[(\mathbb{1}_A - \mathbb{1}_{A^c})X] \leq \mathbb{E}[|X|] \ ,$$

where we used that $|\mathbb{1}_A - \mathbb{1}_{A^c}| = 1$. Hence we conclude (B.7). □

We now state an intermediary lemma, which will be extended to more general assumptions afterwards.

Lemma B.2.3. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and let $Y = \mathbb{E}[X|\mathcal{G}]$. The following assertions hold.*

(i) Equality (B.5) continues to hold with $\mathbb{1}_A$ extended as follows, we have

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all } Z \in L^\infty(\Omega, \mathcal{G}, \mathbb{P}).$$

(ii) For all $Z \in L^\infty(\Omega, \mathcal{G}, \mathbb{P})$, we have

$$\mathbb{E}[XZ|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}] \ .$$

Proof. By linearity of the expectation, (B.5) continues to holds when replacing $\mathbb{1}_A$ by any simple $\mathcal{G}$-measurable random variable Z. Since $X, Y \in L^1$, we also get by dominated convergence that (B.5) continues to holds when replacing $\mathbb{1}_A$ by any bounded $\mathcal{G}$-measurable random variable Z, since we can find Z_n converging pointwise to Z with $|Z_n| \leq |Z|$ for all n. Hence we obtain (i).

Take any $A \in \mathcal{G}$. Then $Z\mathbb{1}_A \in L^\infty(\Omega, \mathcal{G}, \mathbb{P})$ and by applying (i) we get $\mathbb{E}X(Z\mathbb{1}_A) = \mathbb{E}Y(Z\mathbb{1}_A)$. But since $YZ \in L^1(\Omega, \mathcal{G}, \mathbb{P})$, we obtain that $YZ = \mathbb{E}[XZ|\mathcal{G}]$. □

Finally we have the following further extension of identity (B.5).

Lemma B.2.4. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and let $Y = \mathbb{E}[X|\mathcal{G}]$. Equality (B.5) continues to hold with $\mathbb{1}_A$ extended as follows,*

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] \quad \text{for all } \mathcal{G}\text{-measurable r.v. } Z \text{ such that } \mathbb{E}[|XZ|] < \infty.$$

Proof. Take first a non-negative $\mathcal{G}$-measurable r.v. Z such that $\mathbb{E}[|XZ|] < \infty$. Then we can find a non-decreasing sequence $(Z_n)_{n\in\mathbb{N}}$ of simple non-negative $\mathcal{G}$-measurable r.v. Z such that $(Z_n)_{n\in\mathbb{N}}$ converges to Z pointwise. Thus $|XZ_n| = |X|\, Z_n \leq |X|\, Z = |XZ|$, and so by dominated convergence, XZ_n converges to XZ in L^1. Using Lemma B.2.3 (ii), we have that YZ_n is a version of $\mathbb{E}[XZ_n|\mathcal{G}]$ and by Lemma B.2.2, we deduce that $(YZ_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in L^1. Since it converges to YZ pointwise, we conclude that $(YZ_n)_{n\in\mathbb{N}}$ converges to YZ in L^1 by Fatou's lemma. Thus $\mathbb{E}[XZ_n] = \mathbb{E}[YZ_n]$ implies $\mathbb{E}[XZ] = \mathbb{E}[YZ]$ by letting $n \to \infty$. It remains to consider the case of a signed $\mathcal{G}$-measurable r.v. Z such that $\mathbb{E}[|XZ|] < \infty$. In this case, we observe that Z_+, Z_- are non-negative $\mathcal{G}$-measurable r.v.'s such that $\mathbb{E}[|XZ_+|], \mathbb{E}[|XZ_-|] < \infty$. Since $Z = Z_+ - Z_-$ we get the result. □

Many of the useful properties of expectations extend to conditional expectations. We state below some of these useful properties. In the following statements, all equalities and inequalities between random variables, and convergence of such, should be understood to hold $\mathbb{P}$-a.s. The proofs are left as an exercise, (see Exercise B.5).

Proposition B.2.5 (Elementary Properties of Conditional Expectation)**.** *Suppose that* $X, Y, Z, X_n \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ *for all* $n \geq 1$.

*(a) (*linearity*) For all* $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + bY | \mathcal{G}] = a\mathbb{E}[X | \mathcal{G}] + b\mathbb{E}[Y | \mathcal{G}] \ .$$

(b) If X *is* $\mathcal{G}$*-measurable,* $\mathbb{E}[X | \mathcal{G}] = X$.

(c) If $\mathcal{G} = \{\emptyset, \Omega\}$ *is the trivial* σ*-field, then* $\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X]$.

(d) If X *is independent of* $\mathcal{G}$*, then*

$$\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X] \ . \tag{B.8}$$

*(e) (*positivity*) If* $X \leq Y$*, then* $\mathbb{E}[X | \mathcal{G}] \leq \mathbb{E}[Y | \mathcal{G}]$.

(f) $\mathbb{E}[X | \mathcal{G}] \vee \mathbb{E}[Y | \mathcal{G}] \leq \mathbb{E}[X \vee Y | \mathcal{G}]$, $\mathbb{E}[X | \mathcal{G}]_+ \leq \mathbb{E}[X_+ | \mathcal{G}]$ *and* $|\mathbb{E}[X | \mathcal{G}]| \leq \mathbb{E}[|X| | \mathcal{G}]$.

*(g) (*tower property*) If* $\mathcal{H}$ *is a sub-*σ*-field of* $\mathcal{F}$ *such that* $\mathcal{G} \subseteq \mathcal{H}$*, then*

$$\mathbb{E}[\mathbb{E}[X | \mathcal{H}] | \mathcal{G}] = \mathbb{E}[X | \mathcal{G}] \ .$$

(h) The expectation is not modified by conditional expectation,

$$\mathbb{E}[\mathbb{E}[X | \mathcal{G}]] = \mathbb{E}[X] \ .$$

(i) If X *is* $\mathcal{G}$*-measurable and* $XY \in L^1(\Omega, \mathcal{F}, \mathbb{P})$*, then*

$$\mathbb{E}[XY | \mathcal{G}] = X\mathbb{E}[Y | \mathcal{G}] \ . \tag{B.9}$$

(j) If $\sigma(X) \vee \mathcal{H} = \sigma(\sigma(X) \cup \mathcal{H})$ *(the smallest* σ*-field containing* $\sigma(X)$ *and* $\mathcal{H}$*) is independent of* $\mathcal{G}$*, then we have*

$$\mathbb{E}[X | \mathcal{H} \vee \mathcal{G}] = \mathbb{E}[X | \mathcal{H}] \ .$$

We conclude this section with a special case of the conditional expectation.

Definition B.2.2 (Conditional expectation given a random element)**.** *Let* $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$*, and let* X *be a random variable defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *and valued in a measurable space* $(\mathsf{X}, \mathcal{X})$*. Then* $\mathbb{E}[Y | \sigma(X)]$ *is is called the conditional expectation of* Y *given* X*, and equivalently written as* $\mathbb{E}[Y | X]$.

By definition, $\mathbb{E}[Y | X]$ is a $\sigma(X)$-measurable random variable. Thus, by Corollary B.1.7, there exists a Borel function g on X such that $\mathbb{E}[Y | X] = g \circ X$. More precisely we have the following lemma.

Lemma B.2.6. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in a measurable space $(\mathsf{X}, \mathcal{X})$. Then $\mathbb{E}[Y|X] = g(X)$[1], where g is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for all $B \in \mathcal{X}$,*

$$\mathbb{E}[Y\,\mathbb{1}_B(X)] = \mathbb{E}[g(X)\,\mathbb{1}_B(X)]\ . \tag{B.10}$$

Proof. By Corollary B.1.7, there exists a Borel function g on $(\mathsf{X}, \mathcal{X})$ such that $\mathbb{E}[Y|X] = g \circ X$ and any such function g can be chosen provided that, for all $A \in \sigma(X)$

$$\mathbb{E}[Y\,\mathbb{1}_A] = \mathbb{E}[g(X)\,\mathbb{1}_A]\ .$$

Now, observe that $A \in \sigma(X)$ if and only if $A = X^{-1}(B)$ with $B \in \mathcal{X}$. Since $A = X^{-1}(B)$ implies $\mathbb{1}_A = \mathbb{1}_B \circ X$ we get that the previous condition is equivalent to have (B.10) for all $B \in \mathcal{X}$. □

Since $\mathbb{E}[Y|X] = g(X)$ is uniquely defined in $L^1(\Omega, \mathcal{F}, \mathbb{P})$, the choice of g in Lemma B.2.6 is unambiguous in the sense that any two functions g and $\tilde{g}$ satisfying this equality must be equal $\mathbb{P}^X$-a.e.. Also note that by Lemma B.2.6, this function g only depends on $\mathbb{P}^{(X,Y)}$ (which sufficed to compute both sides of the condition (B.10)). In the case where $Y = g \circ Z$ and $X = h \circ Z$, we can thus transport the conditional expectation as we do for the joint distribution. This gives the following lemma.

Lemma B.2.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and Z be a random variable defined on this space and valued in a measurable space $(\mathsf{Z}, \mathcal{Z})$. Let $Y = g \circ Z$ and $X = h \circ Z$, where g and h are measurable functions from $(\mathsf{Z}, \mathcal{Z})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $(\mathsf{X}, \mathcal{X})$, respectively. Suppose that $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, or, equivalently, $g \in L^1\left(\mathsf{Z}, \mathcal{Z}, \mathbb{P}^Z\right)$. Then we have*

$$\mathbb{E}[Y|X] = \mathbb{E}^Z[g \mid h] \circ Z\ ,$$

where $\mathbb{E}^Z[g \mid h]$ denotes the conditional expectation of g given h seen as random variables defined on $\left(\mathsf{Z}, \mathcal{Z}, \mathbb{P}^Z\right)$.

Proof. By definition, $\mathbb{E}[g|h] \in L^1\left(\mathsf{Z}, \mathcal{Z}, \mathbb{P}^Z\right)$ and satisfies, for all $B \in \sigma(h)$

$$\int g\,\mathbb{1}_B\,\mathrm{d}\mathbb{P}^Z = \int \mathbb{E}[g|h]\ \mathbb{1}_B\,\mathrm{d}\mathbb{P}^Z\ ,$$

which can be equivalently written as

$$\mathbb{E}[g(Z)\,\mathbb{1}_B(Z)] = \mathbb{E}[\mathbb{E}[g|h] \circ Z\,\mathbb{1}_B(Z)]\ .$$

We can thus conclude with Lemma B.2.6. □

B.3 Domination: Radon-Nikodym theorem

Let μ, λ be σ-finite measures on $(\Omega, \mathcal{F})$ such that for all $A \in \mathcal{F}$ we have

$$\mu(A) = \int_A \phi\,\mathrm{d}\lambda\ ,$$

where ϕ is some Borel function. Then we say that μ admits a density function ϕ with respect to λ. Note that $\phi \geq 0$ λ-a.s. and the property $\mu(A) = \int_A \phi\,\mathrm{d}\lambda$ for all $A \in \mathcal{F}$ defines ϕ uniquely up to the equality λ-a.e. (See Exercise B.6) Hence the following definition.

[1] It is usual (although not recommended and thus avoided in these lecture notes) to write $\mathbb{E}[Y|X = x]$ for such a $g(x)$.

Definition B.3.1 (Radon-Nikodym derivative)**.** *If $\mu(A) = \int_A \phi \, \mathrm{d}\lambda$ for all $A \in \mathcal{F}$, we say that the λ-*a.e. *equivalent class of ϕ is the Radon-Nikodym derivative of μ with respect to λ and write $\phi = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$.*

We note that if μ admits a density function ϕ with respect to λ, then $\lambda(A) = 0$ implies $\mu(A) = 0$. It turns out that this property is sufficient to have that μ admits a density function with respect to λ.

Definition B.3.2 (Absolute continuity of measures)**.** *Let λ be a measure on $(\Omega, \mathcal{F})$. We say that a σ-finite measure μ is* absolutely continuous with respect to λ *or that λ* dominates *μ and we write $\mu \ll \lambda$ if for all $A \in \mathcal{F}$, $\lambda(A) = 0$ implies $\mu(A) = 0$.*

The following result holds.

Theorem B.3.1 (Radon-Nikodym theorem)**.** *Let $\lambda, \mu \in \mathbb{M}_+(\Omega, \mathcal{F})$ be σ-finite measures on $(\Omega, \mathcal{F})$ such that $\mu \ll \lambda$. Then there exists a non-negative Borel function ϕ such that for all $A \in \mathcal{F}$, $\mu(A) = \int_A \phi \, \mathrm{d}\lambda$.*

Proof. We assume that μ and λ are finite with $\mu \ll \lambda$ (it then easily extends to the case where μ is σ-finite case by partitioning the space with finite μ and λ-measures subsets). Let $\rho = \lambda + \mu$. Let us define, for all $f \in L^2(\Omega, \mathcal{F}, \rho)$,

$$I(f) = \int f \, \mathrm{d}\mu \, .$$

Then we have

$$|I(f)| \leq \int |f| \, \mathrm{d}\mu \leq (\mu(\Omega))^{1/2} \left(\int |f|^2 \, \mathrm{d}\mu \right)^{1/2} \leq (\mu(\Omega))^{1/2} \left(\int |f|^2 \, \mathrm{d}\rho \right)^{1/2} .$$

Hence $f \mapsto I(f)$ is a continuous linear form on $L^2(\Omega, \mathcal{F}, \rho)$. By the Riesz representation theorem (see Theorem A.5.3), there exists $g \in L^2(\Omega, \mathcal{F}, \rho)$ such that for all $f \in L^2(\Omega, \mathcal{F}, \rho)$, $I(f) = \langle f, \bar{g} \rangle$, that is

$$\int f \, \mathrm{d}\mu = \int f \, g \, \mathrm{d}\rho \, . \tag{B.11}$$

Remark moreover that for all $f \geq 0$, $\langle f, \bar{g} \rangle \geq 0$, so one easily gets that the $L^2(\Omega, \mathcal{F}, \rho)$-norm of $\Im(g)$ is zero, hence g is real and the $L^2(\Omega, \mathcal{F}, \rho)$-norm of g_- is zero, hence $g \geq 0$ ρ-a.e.

Let $A \in \mathcal{F}$. We have $\mathbb{1}_A \in L^2(\Omega, \mathcal{F}, \rho)$ and applying (B.11) with $f = \mathbb{1}_A$, we get

$$\mu(A) = \int_A g \, \mathrm{d}\rho \, . \tag{B.12}$$

Take $A = \{g \geq 1\}$. Since $g \mathbb{1}_A \geq \mathbb{1}_A$, we get

$$\mu(A) = \int_A g \, \mathrm{d}\rho \geq \rho(A) = \mu(A) + \lambda(A) \, .$$

Hence $\lambda(A) = 0$ and since $\mu \ll \lambda$ we also have $\mu(A) = 0$ and then $\rho(A) = 0$. Hence finally $0 \leq g < 1$ ρ-a.e. Modifying g by setting it to 0 on the set $\{0 \leq g < 1\}^c$, we thus finally have (B.12) for all $A \in \mathcal{F}$ and $0 \leq g < 1$. Hence since $\rho = \lambda + \mu$, we get that, for all $A \in \mathcal{F}$,

$$\int \mathbb{1}_A (1 - g) \, \mathrm{d}\mu = \int \mathbb{1}_A g \, \mathrm{d}\lambda.$$

This identity extends to any non-negative Borel function f in place of $\mathbb{1}_A$ by monotone convergence and in particular to $f = \mathbb{1}_A/(1-g)$ and obtain

$$\mu(A) = \int_A \phi \, \mathrm{d}\lambda \, ,$$

where $\phi = g/(1-g)$. □

B.4 Conditional Distributions

B.4.1 Regular versions and probability kernels

Definition B.4.1 (Version of Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. We denote*

$$\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A | \, \mathcal{G}] \quad \textit{for any event } A \in \mathcal{F}.$$

A mapping Q on $\Omega \times \mathcal{F}$ valued in $[0,1]$ is called a version of the conditional probability given *$\mathcal{G}$ if, for all $A \in \mathcal{F}$, $\omega \mapsto Q(\omega, A)$ is a version of $A \mapsto \mathbb{P}[A \,|\, \mathcal{G}]$.*

Since $\mathbb{P}[\Omega \,|\, \mathcal{G}] = \mathbb{E}[1 | \, \mathcal{G}] = 1$ and $\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A | \, \mathcal{G}]$ and the conditional expectation $X \mapsto \mathbb{E}[X | \, \mathcal{G}]$ satisfies the usual properties of an expectation (positivity, linearity), we might expect a version Q of the conditional probability given $\mathcal{G}$ to be a (random!) probability on $\mathcal{F}$, in the sense that for $\mathbb{P}$-almost every ω, $A \mapsto Q(\omega, A)$ would be a probability. More precisely we would like to exhibit a *regular conditional probability* as defined in the following.

Definition B.4.2 (Regular Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A* regular version of the conditional probability of $\mathbb{P}$ given *$\mathcal{G}$ is a function*

$$\mathbb{P}^{\mathcal{G}} : \Omega \times \mathcal{F} \to [0,1]$$

such that

(i) For all $A \in \mathcal{F}$, $\mathbb{P}^{\mathcal{G}}(A) : \omega \mapsto \mathbb{P}^{\mathcal{G}}(\omega, A)$ is $\mathcal{G}$-measurable and is a version of the conditional probability of A given $\mathcal{G}$, $\mathbb{P}^{\mathcal{G}}(A)$ is a version of $\mathbb{P}[A \,|\, \mathcal{G}]$, where $\mathbb{P}^{\mathcal{G}}(A)$ denotes the random variable $\mathbb{P}^{\mathcal{G}}(\cdot, A)$;

(ii) For all $\omega \in \Omega$, the mapping $A \mapsto \mathbb{P}^{\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{F}$.

When dealing with a regular conditional probability $\mathbb{P}^{\mathcal{G}}$, one can use all the usual properties of the measure for $\mathbb{P}^{\mathcal{G}}(\omega, \cdot)$. For instance, for any $Y \geq 0$, one can define the random variable $\mathbb{E}^{\mathcal{G}}[Y]$ as

$$\mathbb{E}^{\mathcal{G}}[Y](\omega) = \int Y(\omega') \, \mathbb{P}^{\mathcal{G}}(\omega, \mathrm{d}\omega') \, . \tag{B.13}$$

As usual this definition extends to a signed Y by setting

$$\mathbb{E}^{\mathcal{G}}[Y] = \mathbb{E}^{\mathcal{G}}[Y_+] - \mathbb{E}^{\mathcal{G}}[Y_-] \, ,$$

provided that this difference is well defined $\mathbb{P}$-a.s. (that is, for $\mathbb{P}$-a.e. ω, at least one the terms $\mathbb{E}^{\mathcal{G}}[Y_+]$ or $\mathbb{E}^{\mathcal{G}}[Y_-]$ is finite). All the usual properties (monotone convergence, dominated convergence, Fubini,...) can then be applied for all ω. In particular we have the following result.

Lemma B.4.1. *Let $\mathbb{P}^{\mathcal{G}}$ be a regular version of the conditional probability of $\mathbb{P}$ given $\mathcal{G}$ and let $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then the mapping $\mathbb{E}^{\mathcal{G}}[Y]$ defined by (B.13) is a version of $\mathbb{E}[Y|\mathcal{G}]$.*

Proof. We already have this property for $Y = \mathbb{1}_B$ with $B \in \mathcal{F}$ by Definition B.4.2(i). By linearity of the conditional probability, it remains true for all simple random variables Y. Also by linearity, it is now sufficient to prove the lemma in the case where $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ is non-negative, in which case we can build a non-decreasing sequence of simple random variables $(Y_n)_{n\in\mathbb{N}}$ converging to Y pointwise and so also in L^1. By Lemma B.2.2, it follows that $\mathbb{E}[Y_n|\mathcal{G}] = \mathbb{E}^{\mathcal{G}}[Y_n]$ converges in L^1 to $\mathbb{E}[Y|\mathcal{G}]$. We also know by monotone convergence that $\mathbb{E}^{\mathcal{G}}[Y_n](\omega)$ converges to $\mathbb{E}^{\mathcal{G}}[Y](\omega)$ for every ω, hence $\mathbb{E}^{\mathcal{G}}[Y]$ coincides $\mathbb{P}$-a.s. with the L^1 limit $\mathbb{E}[Y|\mathcal{G}]$ of $\left(\mathbb{E}^{\mathcal{G}}[Y_n]\right)_{n\in\mathbb{N}}$. □

Also, if now Y is a $(\mathsf{Y}, \mathcal{Y})$-valued random variable, for all $A \in \mathcal{Y}$, we have

$$\mathbb{P}^{\mathcal{G}}(Y \in A) = \mathbb{P}[Y \in A \,|\, \mathcal{G}] \ .$$

For each ω the image probability of Y under $\mathbb{P}^{\mathcal{G}}(\omega, \cdot)$ is of course a probability and provides what we call a *regular version of the conditional distribution* of Y given $\mathcal{G}$.

Definition B.4.3 (Regular conditional distribution of Y given $\mathcal{G}$)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space and let Y be an Y-valued random variable. A* regular version of the conditional distribution of Y given $\mathcal{G}$ *is a function*

$$\mathbb{P}^{Y|\mathcal{G}} : \Omega \times \mathcal{Y} \to [0, 1]$$

such that

(i) For all $A \in \mathcal{Y}$, $\omega \mapsto \mathbb{P}^{Y|\mathcal{G}}(\omega, A)$ is $\mathcal{G}$-measurable and is a version of conditional distribution of Y given $\mathcal{G}$, $\mathbb{P}^{Y|\mathcal{G}}(\cdot, A) = \mathbb{P}[Y \in A \,|\, \mathcal{G}]$ $\mathbb{P}$-a.s.

(ii) For every ω, $A \mapsto \mathbb{P}^{Y|\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{Y}$.

Finally, in the case where $\mathcal{G} = \sigma(X)$, the two following definitions are more convenient.

Definition B.4.4. *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A* kernel *Q from $(\mathsf{X}, \mathcal{X})$ onto $(\mathsf{Y}, \mathcal{Y})$ (or, for sake of brevity, on $\mathsf{X} \times \mathcal{Y}$) is a mapping $Q : \mathsf{X} \times \mathcal{Y} \to [0, \infty]$ satisfying the following conditions:*

(i) for every $A \in \mathcal{Y}$, the mapping $Q(\cdot, A) : x \to Q(x, A)$ is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $([0, \infty], \mathcal{B}[0, \infty])$.

(ii) for every $x \in \mathsf{X}$, the mapping $Q(x, \cdot) : A \mapsto Q(x,\ A)$ is a measure on $(\mathsf{Y}, \mathcal{Y})$,

- *Q is called a probability kernel if $Q(x, \mathsf{Y}) = 1$, for all $x \in \mathsf{X}$.*
- *Q is called a Markov kernel if it is a probability kernel on $\mathsf{X} \times \mathcal{X}$, that is, in the case where $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ are the same measurable space.*

Example B.4.1 (Measure seen as a *constant* kernel)**.** *A σ-finite positive measure ν on a space $(\mathsf{Y}, \mathcal{Y})$ can be seen as a kernel on $\mathsf{X} \times \mathcal{Y}$ by defining $N(x, A) = \nu(A)$ for all $x \in \mathsf{X}$ and $A \in \mathcal{Y}$. It is a probability kernel if and only if ν is a probability measure.*

Example B.4.2 (Discrete state-space kernel)**.** *Assume that* X *and* Y *are countable sets. Each element* $x \in \mathsf{X}$ *is then called a state. A kernel* N *on* $\mathsf{X}\times\mathcal{P}(\mathsf{Y})$*, where* $\mathcal{P}(\mathsf{Y})$ *is the set of all parts of* Y*, is specified by a (possibly doubly infinite) matrix* $N = [N(x,y)]_{x,y\in\mathsf{X}\times\mathsf{Y}}$ *with nonnegative entries. Each row* $[N(x,y)]_{y\in\mathsf{Y}}$ *defines a measure on* $(\mathsf{Y},\mathcal{P}(\mathsf{Y}))$ *by setting (using the same notation* N *as for the matrix symbol)*

$$N(x,A) = \sum_{y\in A} N(x,y) \ ,$$

for $A \subset \mathsf{Y}$*. The obtained kernel* N *is a probability kernel if every row* $[N(x,y)]_{y\in\mathsf{Y}}$ *defines a probability on* $(\mathsf{Y},\mathcal{P}(\mathsf{Y}))$*, that is,*

$$\sum_{y\in\mathsf{Y}} N(x,y) = 1$$

for all $x \in \mathsf{X}$*. Note that in the discrete case, the kernel is characterized by setting* $N(x,\{y\}) = N(x,y)$ *for all* $x, y \in \mathsf{X}$*.*

Example B.4.3 (Kernel density function)**.** *Let* λ *be a positive* σ*-finite measure on* $(\mathsf{Y},\mathcal{Y})$ *and* $n : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}_+$ *be a nonnegative function, measurable with respect to the product* σ*-field* $\mathcal{X}\otimes\mathcal{Y}$*. Then, the application* N *defined on* $\mathsf{X}\times\mathcal{Y}$ *by*

$$N(x,A) \ = \int_A n(x,y)\lambda(\mathrm{d}y) \ ,$$

is a kernel. The function n *is called the density function of the kernel* N *with respect to the measure* λ*. The kernel* N *is a probability kernel if and only if* $\int_{\mathsf{Y}} n(x,y)\lambda(\mathrm{d}y) = 1$ *for all* $x \in \mathsf{X}$*.*

Definition B.4.5 (Regular conditional distribution of Y given X)**.** *Let* $(\Omega,\mathcal{F},\mathbb{P})$ *be a probability space and let* X *and* Y *be random variables with values in the measurable spaces* $(\mathsf{X},\mathcal{X})$ *and* $(\mathsf{Y},\mathcal{Y})$*, respectively. A* regular version of the conditional distribution of Y given X *is a probability kernel*

$$\mathbb{P}^{Y|X} : \mathsf{X}\times\mathcal{Y} \to [0,1]$$

such that for all $A \in \mathcal{Y}$*,*

$$\mathbb{P}^{Y|X}(X,A) = \mathbb{P}\left[Y \in A \,|\, X\right] \quad \mathbb{P}\text{-a.s.} \tag{B.14}$$

Going back to the conditional probability $F \to \mathbb{P}\left[F \,|\, \mathcal{G}\right]$, it is not always possible to find a regular version. The difficulty follows from the fact that for each F, $\mathbb{P}\left[F \,|\, \mathcal{G}\right]$ is defined up to a $\mathbb{P}$-null set and this null set may change from one F to another. Because unless in very specific cases the σ-field $\mathcal{F}$ is not countable, there is no guarantee in general that one can choose a particular version such that $F \mapsto \mathbb{P}\left[F \,|\, \mathcal{G}\right]$ is a probability.

However in the context of these lecture notes we can always rely on the following result which says that a regular conditional distribution of Y always exists if Y takes its values in a "nice" space. (We admit this result here, which is actually a very specific case of a more general result where y is valued in a convenient topological spaces, including infinite dimensional ones, see [Dudley, 2002, Theorem 10.2.2]).

Theorem B.4.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let $d \geq 1$ and Y be an $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued random variable. Then there exists a regular version of the conditional distribution of Y given $\mathcal{G}$, $\mathbb{P}^{Y|\mathcal{G}}$, and this version is unique in the sense that for any other regular version $\bar{\mathbb{P}}^{Y|\mathcal{G}}$ of this distribution, for $\mathbb{P}$-almost every ω it holds that*

$$\mathbb{P}^{Y|\mathcal{G}}(\omega, A) = \bar{\mathbb{P}}^{Y|\mathcal{G}}(\omega, A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d) .$$

Moreover, if $\mathcal{G} = \sigma(X)$ for some r.v. X with values in a measurable spaces $(X, \mathcal{X})$, there also exists a unique regular version (hence a probability kernel) $\mathbb{P}^{Y|X}$ of the conditional distribution of Y given X.

When a regular version of a conditional distribution of Y given $\mathcal{G}$ exists, conditional expectations can be written as integrals for each ω, as in Lemma B.4.1. When $\mathcal{G} = \sigma(X)$, it is more convenient to write the integral directly as depending on X, rather than ω, as in the following lemma.

Lemma B.4.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let X and Y be random variables with values in the measurable spaces $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{X}, \mathcal{X})$, respectively, and suppose that $\mathbb{P}^{Y|X}$ is a regular version of the conditional expectation of Y given X. Then for any real-valued measurable function g on Y such that $\mathbb{E}[|g(Y)|] < \infty$, we have*

$$\mathbb{E}[g(Y)|\, X] = \int g(y)\, \mathbb{P}^{Y|X}(X, \mathrm{d}y) \quad \mathbb{P}\text{-a.s.}$$

A very interesting consequence of Theorem B.4.2 is the following result which can be very useful in practice.

Proposition B.4.4. *Let Y and Z be two random vectors valued in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, and F a measurable function from $(\mathbb{R}^{p+q}, \mathcal{B}(\mathbb{R}^{p+q}))$ to $(\mathsf{X}, \mathcal{X})$. Suppose moreover that $\mathbb{E}[|F(Y,Z)|] < \infty$, Y is $\mathcal{G}$-measurable and Z is independent of $\mathcal{G}$. Then we have*

$$\mathbb{E}[F(Y,Z)|\, \mathcal{G}] = \int F(Y,z)\, \mathbb{P}^Z(\mathrm{d}z) . \tag{B.15}$$

(In words, one integrates Z and leaves Y unchanged.)

Proof. By Lemma B.4.1, we only have to determine a regular conditional distribution of $X = (Y,Z)$ given $\mathcal{G}$. Namely, for all $A \in \mathcal{B}(\mathbb{R}^{p+q})$ and all $\omega \in \Omega$, we want to show that we can write

$$\mathbb{P}^{(Y,Z)|\mathcal{G}}(\omega, A) = \mathbb{P}((Y(\omega), Z) \in A) = \int_\Omega \mathbb{1}_A(Y(\omega), Z)\mathbb{P}(\mathrm{d}\tilde{\omega}) . \tag{B.16}$$

Consider first A of the form $A = B \times C$, with $B \in \mathcal{B}(\mathbb{R}^p)$ and $C \in \mathcal{B}(\mathbb{R}^q)$. Then for all $D \in \mathcal{G}$, we have since $\mathbb{1}_D Y$ is $\mathcal{G}$ measurable and Z is independent of $\mathcal{G}$

$$\begin{aligned}\mathbb{E}[\mathbb{1}_D \mathbb{1}_A(Y,Z)] = \mathbb{E}[\mathbb{1}_D \mathbb{1}_B(Y)\mathbb{1}_C(Z)] &= \mathbb{E}[\mathbb{1}_D \mathbb{1}_B(Y)]\, \mathbb{E}[\mathbb{1}_C(Z)] \\ &= \mathbb{E}[\mathbb{1}_D \mathbb{1}_B(Y)]\, \mathbb{P}(Z \in C) = \mathbb{E}[\mathbb{1}_D \mathbb{1}_B(Y)\mathbb{P}(Z \in C)] .\end{aligned}$$

Therefore, we have almost surely

$$\mathbb{P}^{(Y,Z)|\mathcal{G}}(\omega, A) = \mathbb{E}[\mathbb{1}_A(Y,Z)|\, \mathcal{G}](\omega) = \mathbb{1}_B(Y(\omega))\mathbb{P}(Z \in C) = \int_\Omega \mathbb{1}_A(Y(\omega), Z(\tilde{\omega}))\mathbb{P}(\mathrm{d}\tilde{\omega}) . \tag{B.17}$$

Consider now the two set $\mathcal{E}$ and $\mathcal{C}$ contained in $\mathcal{F} = \mathcal{B}(\mathbb{R}^{p+q})$ defined by

$$\mathcal{E} = \left\{ A \in \mathcal{F} \,:\, \mathbb{P}^{(Y,Z)|\mathcal{G}}(\omega, A) = \int_\Omega \mathbb{1}_A(Y(\omega), Z(\tilde{\omega}))\mathbb{P}(\mathrm{d}\tilde{\omega}) \text{ , } \omega\text{-almost surely} \right\}$$

$$\mathcal{C} = \{ B \cap C \,:\, B \in \mathcal{B}(\mathbb{R}^p) \text{ and } C \in \mathcal{B}(\mathbb{R}^q) \} \,.$$

By (B.17), we get that $\mathcal{C} \subset \mathcal{E}$. It is straightforward to check that $\mathcal{C}$ is stable by finite intersection, contains Ω and $\sigma(\mathcal{C}) = \mathcal{H} \vee \mathcal{G}$. Therefore it is a π-system. Then we just need to show that $\mathcal{E}$ is a λ-system since by the π-λ theorem, it will imply that $\sigma(\mathcal{C}) = \mathcal{H} \vee \mathcal{G} \subset \mathcal{E}$.

Let $A \in \mathcal{E}$, A it is clear by definition that $A^{\mathrm{c}} \in \mathcal{E}$. Consider now a sequence $(A_n)_{n\in\mathbb{N}} \in \mathcal{C}^{\mathbb{N}}$ such that for all $n < p$, $A_n \cap A_p = \emptyset$. Then by defition for all $N \in \mathbb{N}$, we have almost surely

$$\mathbb{P}^{(Y,Z)|\mathcal{G}}\left(\omega, \bigcup_{k=0}^{N} A_k\right) = \int_\Omega \mathbb{1}_{\bigcup_{k=0}^N A_k}(Y(\omega), Z(\tilde{\omega}))\mathbb{P}(\mathrm{d}\tilde{\omega}) \,. \tag{B.18}$$

Therefore almost surely for all $N \in \mathbb{N}$ (note the difference here), we get that (B.18) holds. Setting $A = \cup_{k=0}^N A_k$ and using the monotone convergence theorem, we get

$$\mathbb{P}^{(Y,Z)|\mathcal{G}}(\omega, A) = \int_\Omega \mathbb{1}_A(Y(\omega), Z(\tilde{\omega}))\mathbb{P}(\mathrm{d}\tilde{\omega}) \,. \tag{B.19}$$

Then $A \in \mathcal{E}$ and $\mathcal{E}$ is a λ-system. So we have shown (B.16). □

B.4.2 Disintegration of a measure on a product space

Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. Recall that, given two σ-finite measures $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ and $\nu \in \mathbb{M}_+(\mathsf{Y}, \mathcal{Y})$, the product measure $\mu \otimes \nu$ is uniquely defined on $(\mathsf{X}\times\mathsf{Y}, \mathcal{X}\otimes\mathcal{Y})$ by

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B) \,, \qquad A \in \mathcal{X},\, B \in \mathcal{Y} \,. \tag{B.20}$$

Moreover, in the case of probability measures, (X, Y) has distribution $\mu \otimes \nu$ exactly means that X has distribution μ, Y has distribution ν and X and Y are independent. This can be summarized as

$$\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes \mathbb{P}^Y \quad \text{with } \mathbb{P}^X = \mu \text{ and } \mathbb{P}^Y = \nu \,. \tag{B.21}$$

Now consider any two random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. The existence of a regular version $\mathbb{P}^{Y|X}$ of the conditional distribution of Y given X allows us to extend the above product, which only holds if X and Y are independent to the following one, which will always hold :

$$\mathbb{P}^{(X,Y)} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X} \,. \tag{B.22}$$

This is called a disintegration of $\mathbb{P}^{(X,Y)}$. The goal of this section is to define the $\otimes$ that appears in this formula, which can be seen as an extension of the identity

$$\mathbb{P}^{(X,Y)}(A \times B) = \mathbb{E}\left[\mathbb{1}_A(X)\mathbb{E}\left[\mathbb{1}_B(Y)\middle|\, X\right]\right] = \mathbb{E}\left[\mathbb{1}_A(X)\mathbb{P}^{Y|X}(X, B)\right] \,, \tag{B.23}$$

exactly as (B.21) is an extension of (B.20).

Let us first explain how general kernels also act on measures. Let N be a kernel on $\mathsf{X}\times\mathcal{Y}$ and μ be a positive measure on $(\mathsf{X}, \mathcal{X})$ and define the measure μN by

$$\mu N(A) = \int_{\mathsf{Y}} \mu(\mathrm{d}x)\ N(x,\ A) \,, \qquad A \in \mathcal{Y} \,.$$

Proposition B.4.5. *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. Let N be a kernel on $\mathsf{X} \times \mathcal{Y}$ and $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Then $\mu N \in \mathbb{M}_+(\mathsf{Y}, \mathcal{Y})$. If N is a probability kernel, then $\mu N(\mathsf{X}) = \mu(\mathsf{X})$. Hence, if moreover $\mu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$, then $\mu N \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$.*

Proof. Note first that $\mu N(A) \geq 0$ for all $A \in \mathcal{Y}$ and $\mu N(\emptyset) = 0$ since $N(x, \emptyset) = 0$ for all $x \in \mathsf{X}$. Therefore, it suffices to establish the countable additivity of μN. Let $(A_i)_{i \in \mathbb{N}} \subset \mathcal{X}$ be a sequence of pairwise disjoints sets. Since $N(x, \cdot)$ is a measure for all $x \in \mathsf{X}$, the countable additivity implies that $N(x, \bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} N(x, A_i)$. Moreover, the function $x \mapsto N(x, A_i)$ is nonnegative and measurable for all $i \in \mathbb{N}$, thus the monotone convergence theorem yields

$$\mu N\left(\bigcup_{i=1}^{\infty} A_i\right) = \int N\left(x, \bigcup_{i=1}^{\infty} A_i\right) \mu(\mathrm{d}x) = \int \sum_{i=1}^{\infty} N(x, A_i) \, \mu(\mathrm{d}x) = \sum_{i=1}^{\infty} \mu N(A_i) \; .$$

□

Next we introduce the notation $\mu \otimes N$.

Theorem B.4.6. *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. Let N be a probability kernel on $\mathsf{X} \times \mathcal{Y}$ and $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$. Then there exists a unique measure $\mu \otimes N$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$, such that, for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, we have*

$$\mu \otimes N(A \times B) = \int \mathbb{1}_A(x) \, N(x, B) \, \nu(\mathrm{d}x) \; . \tag{B.24}$$

Moreover, for all $f \in \mathrm{F}_+(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$, the following assertions hold.

(i) for all $x \in \mathsf{X}$, $y \mapsto f(x, y)$ belongs to $\mathrm{F}_+(\mathsf{Y}, \mathcal{Y})$,

(ii) $x \mapsto \int f(x, y) \, N(x, \mathrm{d}y)$ belongs to $\mathrm{F}_+(\mathsf{X}, \mathcal{X})$

(iii) $$\int f \, \mathrm{d}\mu \otimes N = \int \left(\int f(x, y) \, N(x, \mathrm{d}y) \right) \mu(\mathrm{d}x) \; .$$

Proof. This result clearly extends the usual Tonelli theorem that allows one to define product of measures, and the proof is similar and follows the following steps.

Step 1 We first show Assertions (i) and (ii) when $f = \mathbb{1}_C$ with $C \in \mathcal{X} \otimes \mathcal{Y}$. This follows from Theorem B.1.1 by checking that

$$\{C \in \mathcal{X} \otimes \mathcal{Y} \, : \, \text{Assertions (i) and (ii) hold with } f = \mathbb{1}_C\}$$

is a λ-system that contains the π-system $\{A \times B \, : \, A \in \mathcal{X}, \, B \in \mathcal{Y}\}$.

Step 2 We next extend Assertions (i) and (ii) to any $f \in \mathrm{F}_+(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. This follows from the approximation of f given by Proposition B.1.6 and, for assertion (ii) from the monotonous convergence theorem.

Step 3 Now we prove the existence of a measure $\mu \otimes N$ satisfying (B.20). Using Assertion (i) and the monotonous convergence theorem, for all $x \in \mathsf{X}$, since $N(x, \cdot)$ is a measure, $C \mapsto \int \mathbb{1}_C(x, y) \, N(x, \mathrm{d}y)$ is σ-additive on $\mathcal{X} \otimes \mathcal{Y}$. Using the monotonous convergence theorem again, we obtain that $C \mapsto \int \left(\int \mathbb{1}_C(x, y) \, N(x, \mathrm{d}y) \right) \mu(\mathrm{d}x)$ also is σ-additive on $\mathcal{X} \otimes \mathcal{Y}$. Thus $C \mapsto \int \left(\int \mathbb{1}_C(x, y) \, N(x, \mathrm{d}y) \right) \mu(\mathrm{d}x)$ defines a measure on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ that coincide with the right-hand side of (B.24) when $C = A \times B$.

Step 4 When $\mu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$ the uniqueness for the measure $\mu \otimes N$ directly follows from Theorem B.1.5 and the fact that $\{A \times B \,:\, A \in \mathcal{X},\, B \in \mathcal{Y}\}$ is a π-system. The more general σ-finite assumption $\mu \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ easily follows by partitioning the space with μ-finite subsets.

Step 5 Finally, as in the first two steps, we can check Assertion (iii) when $f = \mathbb{1}_C$ with $C \in \mathcal{X} \otimes \mathcal{Y}$ using Theorem B.1.1 with the λ-system

$$\{C \in \mathcal{X} \otimes \mathcal{Y} \,:\, \text{Assertion (iii) holds with } f = \mathbb{1}_C\}$$

that contains the π-system $\{A \times B \,:\, A \in \mathcal{X},\, B \in \mathcal{Y}\}$ by (B.24). As usual, Assertion (iii) extends to all $f \in \mathrm{F}_+(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ by the monotonous convergence theorem and Proposition B.1.6.

□

It is immediate to see that if, in Theorem B.4.6, $\mu \in \mathbb{M}_1(\mathsf{X}, \mathcal{X})$, then $\mu \otimes N \in \mathbb{M}_1(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Applying Theorem B.4.6, we in particular have that, for two random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, if $\mathbb{P}^{Y|X}$ is well defined, we have, for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{P}^X \otimes \mathbb{P}^{Y|X}(A \times B) &= \int \left(\int \mathbb{1}_A(x) \mathbb{1}_B(y)\, \mathbb{P}^{Y|X}(x, \mathrm{d}y) \right)\, \mathbb{P}^X(\mathrm{d}x) \\
&= \int \mathbb{1}_A(x) \left(\int \mathbb{1}_B(y)\, \mathbb{P}^{Y|X}(x, \mathrm{d}y) \right)\, \mathbb{P}^X(\mathrm{d}x) \\
&= \mathbb{E}\left[\mathbb{1}_A(X) \mathbb{P}^{Y|X}(X, B) \right] \\
&= \mathbb{P}^{(X,Y)}\left(A \times B\right) \,.
\end{aligned}$$

The last equality were already noted in (B.23). So with Theorem B.4.6, we have indeed that the *disintegration formula* (B.22) holds as soon as $\mathbb{P}^{Y|X}$ is well defined. The converse is also true: if one is able to "disintegrate" $\mathbb{P}^{(X,Y)}$ as a $\otimes$-product between a probability measure and a probability kernel, one can identify the probability as $\mathbb{P}^X$ and the kernel as (a regular version of) $\mathbb{P}^{Y|X}$. We state this as follows.

Theorem B.4.7. *Let X and Y be two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. Let μ be a probability on $(\mathsf{X}, \mathcal{X})$ and N be a probability kernel on $\mathsf{X} \times \mathcal{Y}$. Then we have $\mathbb{P}^{(X,Y)} = \mu \otimes N$ if and only if $\mathbb{P}^X = \mu$ and N is a regular version of the conditional distribution of Y given X. Moreover it then follows that $\mathbb{P}^Y = \mu N$.*

Proof. As explained above, the "if" part follows from observing that (B.22) holds as soon as a regular version of the conditional distribution of Y given X exists, as a consequence of (B.23). Now, the "only if" part in turn follows from the converse implication: if $\mathbb{P}^{(X,Y)} = \mu \otimes N$ then, for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

$$\mathbb{P}^{(X,Y)}(A \times B) = \int \mathbb{1}_A(x) N(x, B)\, \mu(\mathrm{d}x) \,.$$

In particular taking $B = \mathsf{Y}$ yields $\mu = \mathbb{P}^X$ but then the previously displayed equation can be written as

$$\mathbb{P}^{(X,Y)}(A \times B) = \mathbb{E}\left[\mathbb{1}_A(X) N(X, B) \right] \,,$$

which yields $N(X, B) = \mathbb{E}\left[\mathbb{1}_B(Y) \middle| X \right]$, from which we conclude that $\mathbb{P}^{Y|X} = N$. □

B.4.3 Conditional distribution for Gaussian vectors

An important application of the projection theorem in Hilbert spaces is the computation of the conditional mean for L^2 random variables, see Lemma B.2.1(ii). It also provides an easy way to compute the conditional distribution in a Gaussian context, where the following result holds. The poof is left as an exercise (see Exercise B.7).

Proposition B.4.8. *Let $p, q \geq 1$. Let $\mathbf{X}$ and $\mathbf{Y}$ be two jointly Gaussian vectors, respectively valued in $\mathbb{R}^p$ and $\mathbb{R}^q$. Define*

$$\widehat{\mathbf{X}} := \mathrm{proj}\left(\mathbf{X} | \left\{a + B\mathbf{Y} \,:\, a \in \mathbb{R}^p\,,\; B \in \mathbb{R}^{p\times q}\right\}\right)\,.$$

Then the following assertions hold.

(i) We have

$$\mathbb{E}\left[\mathbf{X} | \mathbf{Y}\right] = \widehat{\mathbf{X}}\,.$$

(ii) We have

$$\mathrm{Cov}(\mathbf{X} - \widehat{\mathbf{X}}) = \mathbb{E}\left[\mathbf{X}(\mathbf{X} - \widehat{\mathbf{X}})^T\right] = \mathbb{E}\left[(\mathbf{X} - \widehat{\mathbf{X}})\mathbf{X}^T\right]$$

and the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ is given by

$$\mathbb{P}^{\mathbf{X}|\mathbf{Y}}(\mathbf{Y}, \cdot) = \mathcal{N}(\widehat{\mathbf{X}}, \mathrm{Cov}(\mathbf{X} - \widehat{\mathbf{X}}))\,.$$

(iii) If $\mathrm{Cov}(\mathbf{Y})$ is invertible, then $\widehat{\mathbf{X}}$ is given by

$$\widehat{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\right] + \mathrm{Cov}\left(\mathbf{X}, \mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\left(\mathbf{Y} - \mathbb{E}\left[\mathbf{Y}\right]\right),$$

and

$$\mathrm{Cov}(\mathbf{X} - \widehat{\mathbf{X}}) = \mathrm{Cov}(\mathbf{X}) - \mathrm{Cov}\left(\mathbf{X}, \mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\mathrm{Cov}\left(\mathbf{Y}, \mathbf{X}\right)\,.$$

B.4.4 Conditional density function

The following definition is classically seen in elementary probability courses in the discrete case (ξ and ξ' are compting measures on countable sets) or in the Lebesgue case (ξ and ξ' are Lebesgue measures of given dimensions).

Definition B.4.6 (Conditional density function)**.** *Let (X, Y) be two random elements admitting a density function f with respect to $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then the function $(x, y) \mapsto f(y|x)$ defined for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ by*

$$f(y|x) = \frac{f(x, y)}{\int f(x, y')\,\mathrm{d}\xi'(y')}$$

is called the conditional density function of Y given X.

Observe that the denominator is the density function of X with respect to ξ applied to x. Hence (see Exercise B.6) it satisfies

$$0 < \int f(x, y')\,\mathrm{d}\xi'(y') < \infty \quad \text{for } \mathbb{P}^X\text{-a.e. } x.$$

For x's such that this is not true, we can set $y \mapsto f(y|x)$ to be any arbitrary density function such as the density function of Y, in which case

$$y \mapsto f(y|x) \text{ is a density function with respect to } \xi' \text{ for all } x \in \mathsf{X}.$$

In the case where (X, Y) satisfies the assumptions of Definition B.4.6, the conditional distribution of Y given X is given by the following result, whose proof is left as an exercise (Exercise B.8).

Theorem B.4.9. *Let (X, Y) be a random pair admitting a density function $f : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ with respect to $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then the conditional distribution of Y given X is given by*

$$\mathbb{P}^{Y|X}(x, A) = \int_A f(y|x)\, \xi'(\mathrm{d}y) \quad \textit{for all } x \in \mathsf{X} \textit{ and } A \in \mathcal{Y}\,,$$

where $(x, y) \mapsto f(y|x)$ is the conditional density function of Y given X.

B.5 Exercises

Exercise B.1. Prove Lemma B.1.2. [*Hint:* Note that if $A \subseteq B$ are two subsets of Ω, then $A \cap B^c = \emptyset$ and $B \setminus A = (A \cup B^c)^c$.]

Exercise B.2. Prove Lemma B.1.3. [*Hint:* Note that $A^c \cap B = B \setminus (A \cap B)$ and use Lemma B.1.2.]

Exercise B.3 (Stochastic order). Let F be the distribution function of a probability on $\mathbb{R}$. Define (as usual) its *pseudo-inverse* by

$$F^{-1}(t) = \inf\{x :\ F(x) \geq t\}\,, \qquad t \in (0, 1)\,.$$

1. Show that for all $t \in (0, 1)$, $F \circ F^{-1}(t) \geq t$, with equality if F is continuous.
2. Show that for all $x \in \mathbb{R}$, $F^{-1} \circ F(x) \leq x$, , with equality if F is (strictly) increasing.
3. Let X be a r.v. with distribution function F; show that $X = F^{-1}(F(X))$ a.s.
4. Show that if F is continuous, then $F(X)$ has a uniform distribution on $[0, 1]$. Compute $\mathbb{E}\left[F^n(X)\right]$ for all $n \in \mathbb{N}$.
5. Show that if U is a uniform r.v. on $[0, 1]$, then $F^{-1}(U)$ has distribution function F.

Let X and Y two r.v. with distribution functions F and G. Suppose that for all $s \in \mathbb{R}$, $F(s) \leq G(s)$. We say that F is stochastically larger than or equal to G and we denote $G \leq_{sto} F$.

6. Provide examples of F and G such that $G \leq_{sto} F$.
7. Show that $G \leq_{sto} F$ if and only if there exists a random bidimensional vector (X, Y) such that $X \sim F$, $Y \sim G$ (this is called a *coupling* with marginals (F, G)) and $Y \leq X$ a.s.
8. Provide an example of F and G such that $G \leq_{sto} F$ and a coupling (X, Y) with marginals (F, G) for which $\mathbb{P}(Y \leq X) < 1$.
9. Show that if $G \leq_{sto} F$, then for any non-decreasing $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[f(Y)] \leq \mathbb{E}[f(X)]$ for any coupling (X, Y) with marginals (F, G).

Exercise B.4. Let Y be an L^2 real valued r.v. defined on $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\mathcal{B}$ be a sub-σ-field of $\mathcal{A}$. Define the conditional variance of Y given $\mathcal{B}$ by

$$\sigma^2(Y|\mathcal{B}) = \mathbb{E}[Y^2 \mid \mathcal{B}] - (\mathbb{E}[Y \mid \mathcal{B}])^2.$$

Denoting by $\sigma^2(Z)$ the variance of Z, show that

$$\sigma^2(Y) = \sigma^2(\mathbb{E}[Y \mid \mathcal{B}]) + \mathbb{E}[\sigma^2(Y|\mathcal{B})].$$

What is the result when Y is independent of $\mathcal{B}$?

Exercise B.5. Show all the properties of Proposition B.2.5. [*Hint for Assertion (j)* : first take an element of $\mathcal{H} \vee \mathcal{G}$ of the form $A \cap B$. Use the $\pi - \lambda$-theorem to conclude.]

Exercise B.6. Let μ and λ be two σ-finite measures on $(\Omega, \mathcal{F})$ and let a Borel function $\phi : \Omega \to \bar{\mathbb{R}}_+$ satisfy, for all $A \in \mathcal{F}$,

$$\mu(A) = \int_A \phi \, \mathrm{d}\lambda \, . \tag{B.25}$$

1. Show that if a Borel function $\psi : \Omega \to \bar{\mathbb{R}}$ satisfies, for all $A \in \mathcal{F}$,

   $$\int_A \psi \, \mathrm{d}\lambda = 0 \, ,$$

   then $\psi = 0$ λ-a.e.

2. Show that , for all $A \in \mathcal{F}$ such that $\lambda(A) = 0$, we have $\mu(A) = 0$.

3. Show that $\phi > 0$ μ-a.e.

4. Give an example where we do not have that $\phi > 0$ λ-a.e.

5. Show that $\phi < \infty$ λ-a.e. [Hint : first consider the case where μ is finite]

6. Show that (B.25) uniquely defines ϕ up to a λ-null set.

Exercise B.7. Let $\mathbf{X}$ and $\mathbf{Y}$ be as in Proposition B.4.8.

1. In order to prove Proposition B.4.8(i) and (ii), use Proposition B.4.4.

2. Use the characterization of the orthogonal projection to prove Proposition B.4.8(iii).

Exercise B.8. Prove Theorem B.4.9.

Exercise B.9. Let $(\mathbf{X}, \mathbf{Y})$ be a random vector valued in $\mathbb{R}^{p+q}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Show that the conditional density of $\mathbf{X}$ given $\mathbf{Y}$ is always well defined when

1. $\mathbf{X}$ and $\mathbf{Y}$ are discrete random variables (i.e. take values in countable sets).

2. $(\mathbf{X}, \mathbf{Y})$ admit a density with respect to the Lebesgue measure.

3. $\mathbf{Y}$ is a discrete random variable [Hint : show that $(\mathbf{X}, \mathbf{Y})$ admits a density with respect to $\xi \otimes \xi'$ with ξ' a counting measure and $\xi = \mathbb{P}^{\mathbf{X}}$.].

4. Deduce a formula for $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ in all the previous cases.

5. Give a new proof of Proposition B.4.8(i) and (ii) in the case where $(\mathbf{X}, \mathbf{Y})$ has an invertible covariance matrix.

Exercise B.10. Let X be a real valued random variable admitting a symmetric density f (with respect to the Lebesgue measure), $f(-x) = f(x)$ for all $x \in \mathbb{R}$.

1. Compute the conditional distribution of X given $|X|$.
2. Let Y be any other random variable defined on the same probability space as X and let $\mathbb{P}^{X|Y}$ denote the conditional distribution of X given Y. Show that for all Borel set A with null Lebesgue measure, we have that, for $\mathbb{P}^Y$-a.e. y, $\mathbb{P}^{X|Y}(y, A) = 0$.
3. Do we have that $\mathbb{P}^{X|Y}(y, \cdot)$ admits a density (with respect to the Lebesgue measure) for $\mathbb{P}^Y$-a.e. y?

Exercise B.11. Let (X, Y) be an $\mathbb{R}^2$-valued r.v. Determine the conditional expectation and distribution of X given Y in the following choices for the distribution of (X, Y) :

1. the uniform distribution on the triangle $(0, 0), (1, 0), (0, 1)$
2. the uniform distribution on the square $(0, 0), (1, 0), (1, 1), (0, 1)$

Exercise B.12. Let X and Y be two r.v.s defined on the same probability space. Suppose that X has its values in $\mathbb{N}$ and Y follows an exponential distribution with unit mean. Suppose also that the conditional distribution of X given Y is Poisson with mean Y. Determine the distribution of (X, Y) and that of X. Compute the conditional distribution of Y given X.

Exercise B.13. Let $X_1, \ldots, X_p$ be independent r.v.'s following Poisson distributions with parameters $\lambda_1, \ldots, \lambda_p$.

1. Determine the conditional distribution of $(X_1, \ldots, X_{p-1})$ given $X_1 + \ldots + X_p$.
2. Compute $\mathbb{E}[X_1 \mid X_1 + X_2]$.

Exercise B.14. Let $X_1, \ldots, X_n$ be i.i.d. random variables with density f, assumed to be continuous on $\mathbb{R}$.

1. Recall that the order statistic $(X_{(1)}, \ldots, X_{(n)})$ obtained by ordering$X_1, \ldots, X_n$ in an increasing order admits a density.
2. Determine the conditional distribution of $\min_{1 \le i \le n} X_i$ given $\max_{1 \le i \le n} X_i$.
3. Assuming that $\mathbb{E}[|X_i|] < \infty$ for all $i = 1, \ldots, n$, deduce an expression of $\mathbb{E}[\min_{1 \le i \le n} X_i \mid \max_{1 \le i \le n} X_i]$.

Exercise B.15 (Order statistic). Let $X = (X_1, \ldots, X_n)$ be a random vector with density $f(x)$, $x \in \mathbb{R}^n$. Let $R = (R(1), \ldots, R(n))$ be *rank statistic* of X, that is, for all $i \in \{1, \ldots, n\}$,

$$R(i) = \sum_{j=1}^{n} \mathbb{1}_{\{X_i \ge X_j\}} \, .$$

1. Show that, with probability 1, there exists a permutation σ of $\{1, \ldots, n\}$ such that $X_{\sigma(1)} < \ldots < X_{\sigma(n)}$. What is the relationship between σ and R ?

2. Show that, for any permutation r of $\{1,\dots,n\}$ and all borel function $g:\mathbb{R}^n\to\mathbb{R}_+$

$$\mathbb{E}\left[g(X_{\sigma(1)},\dots,X_{\sigma(n)})\mathbb{1}_{\{R=r\}}\right]=\int g(x_1,\dots,x_n)\mathbb{1}_{\{x_1<\dots<x_n\}}f(x_{r(1)},\dots,x_{r(n)})\,\mathrm{d}x_1\dots\mathrm{d}x_n$$

3. Deduce the conditional distribution of R given $(X_{\sigma(1)},\dots,X_{\sigma(n)})$.

4. What do you obtain when the X_i's are iid ?

Appendix C

Convergence of random elements in a metric space

In this appendix we provide the main definitions and results concerning the convergence of a sequence of random elements valued in a metric space. The strong convergence and the convergence in probability are not more difficult in this setting than in the case of vector valued random variables. The weak convergence is more delicate as some topology properties of the metric space have to be considered. A classical reference for the weak convergence in metric spaces is Billingsley [1999a]. Here we provide a brief account of the essential classical definitions and results. The detailed proofs can be found in Billingsley [1999a].

From now on, we let (X, d) be a metric space. We note $\mathrm{C}_b(\mathsf{X})$ (resp. $\mathrm{Lip}_b(\mathsf{X})$) the space of real-valued bounded continuous functions (resp. bounded and Lipschitz) on (X, d). We denote by $\mathcal{B}(\mathsf{X})$ the Borel σ-fields on X and by $\mathbb{M}_1(\mathsf{X})$ the set of probability measures on $\mathcal{B}(\mathsf{X})$.

C.1 Definitions and characterizations

As mentioned above, the weak convergence is in general more delicate to handle than other convergences. An additional difficulty is that it is often presented as a "convergence" of a sequence of random variables, but the word "convergence" is not rigorous in such a presentation. In fact the weak convergence defines a convergence for the sequence of the marginal distributions, thus, for a sequence valued in $\mathbb{M}_1(\mathsf{X})$, the set of probability measures on X.

The term weak convergence is opposed to strong convergence which, in contrast, makes sense only for a sequence of random variables.

Definition C.1.1 (Strong convergence)**.** *Let X, X_n, $n \geq 1$ be random variables valued in $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will say that X_n strongly converges to X and denote $X_n \xrightarrow{\text{a.s.}} X$ in (X, d) (or simply $X_n \xrightarrow{\text{a.s.}} X$ if no ambiguity occurs) if $d(X_n, X) \to 0$ almost surely.*

A basic criterion for proving strong convergence is based on the Borel Cantelli lemma.

Lemma C.1.1 (Borel Cantelli's Lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of measurable sets. Then,*

$$\sum_{k\in\mathbb{N}} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}(\limsup A_n) = 0 \,.$$

In particular, if X, X_n, $n \geq 1$ *are random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that, for any* $\epsilon > 0$,

$$\sum_{k \in \mathbb{N}} \mathbb{P}(d(X_n, X) > \epsilon) < \infty \ ,$$

then $X_n \xrightarrow{\text{a.s.}} X$.

The convergence in probability also applies to a sequence of random variables. It is weaker than the strong convergence.

Definition C.1.2 (Convergence in probability)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *We will say that* X_n *converges in probability to* X *and denote* $X_n \xrightarrow{P} X$ *in* (X, d) *(or simply* $X_n \xrightarrow{P} X$ *if no ambiguity occurs) if* $\mathbb{P}(d(X_n, X) > \epsilon) \to 0$ *for any* $\epsilon > 0$.

It is straightforward to verify that the convergence in probability can be characterized with the strong convergence as follows.

Theorem C.1.2. *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Then we have* $X_n \xrightarrow{P} X$ *if and only if for all subsequence* (X_{α_n}), *there is a subsequence* $(X_{\alpha_{\beta_n}})$ *such that* $X_{\alpha_{\beta_n}} \xrightarrow{\text{a.s.}} X$.

The following result shows that any probability measure μ defined on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ is *regular*, in the sense that it can be defined for all $A \in \mathcal{B}(\mathsf{X})$ by

$$\mu(A) = \inf \{\mu(U) \ : \ U \text{ open set } \supset A\} = \sup \{\mu(F) \ : \ F \text{ closed set} \subset A\} \ . \tag{C.1}$$

Proposition C.1.3. *Let* $\mu \in \mathbb{M}_1(\mathsf{X})$. *Then (C.1) holds for all* $A \in \mathcal{B}(\mathsf{X})$.

Definition C.1.3 (Weak convergence of probability measures)**.** *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$. *We say that* μ_n *converges weakly to* μ *if, for all* $f \in \mathrm{C}_b(\mathsf{X})$, $\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$.

Weak convergence is also often used for a sequence of random variables.

Definition C.1.4 (Weak convergence of random variables)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. *We will say that* X_n *converges weakly to* X *and denote* $X_n \Longrightarrow X$ *in* (X, d) *(or simply* $X_n \Longrightarrow X$ *if no ambiguity occurs) if* μ_n *converges weakly to* μ, *where* μ_n *is the probability distribution of* X_n *and* μ *is the probability distribution of* X.

The following theorem provides various characterizations of the weak convergence. It is often referred to as the *Portmanteau theorem*.

Theorem C.1.4. *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$. *The following properties are equivalent:*

(i) μ_n *converges weakly to* μ,

(ii) for all $f \in \mathrm{Lip}_b(\mathsf{X})$, $\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$,

(iii) for all closed set F, $\limsup_n \mu_n(F) \leq \mu(F)$,

(iv) for all open set U, $\liminf_n \mu_n(U) \geq \mu(U)$,

(v) for all $B \in \mathcal{B}(\mathsf{X})$ *such that* $\mu(\partial B) = 0$, $\lim_n \mu_n(B) = \mu(B)$.

Let (Y, δ) be a metric space. For all measurable $h : \mathsf{X} \to \mathsf{Y}$, we denote

$$D_h \stackrel{\text{def}}{=} \{x \in \mathsf{X} : \ h \text{ is discontinuous at } x\} \ . \tag{C.2}$$

The following theorem is often referred to as the *continuous mapping theorem.*

Theorem C.1.5. *Let* $\mu_n, \mu \in \mathbb{M}_1(\mathsf{X})$ *and* $h : \mathsf{X} \to \mathsf{Y}$ *be measurable. We assume that* μ_n *converges weakly to* μ *and that* $\mu(D_h) = 0$. *Then* $\mu_n \circ h^{-1}$ *converges weakly to* $\mu \circ h^{-1}$.

The weak convergence is equivalent to the convergence of integrals of bounded continuous functions. The case of unbounded continuous functions is treated in the following result.

Proposition C.1.6. *Assume that* μ_n *converges weakly to* μ. *Let* f *be a continuous function such that* $\lim_{a\to\infty} \sup_n \int_{|f|>a} |f| \, \mathrm{d}\mu_n = 0$. *Then* f *is* μ*-integrable and* $\int f d\mu_n \to \int f d\mu$.

We now provide a statement expressed with random variables for convenience and add the equivalent statement for the strong convergence and the convergence in probability. It is a direct application of Theorem C.1.5 and Theorem C.1.2.

Theorem C.1.7 (Continuous mapping theorem for the three convergences)**.** *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $h : \mathsf{X} \to \mathsf{Y}$ *measurable and define* D_h *as in (C.2). Assume that* $\mathbb{P}(X \in D_h) = 0$. *Then the following assertions hold.*

*(*i*) If* $X_n \xrightarrow{\text{a.s.}} X$, *then* $h(X_n) \xrightarrow{\text{a.s.}} h(X)$.

*(*ii*) If* $X_n \xrightarrow{P} X$, *then* $h(X_n) \xrightarrow{P} h(X)$.

*(*iii*) If* $X_n \Longrightarrow X$, *then* $h(X_n) \Longrightarrow h(X)$.

Let us recall briefly some standard results on the weak convergence, strong convergence and convergence in probability.

Theorem C.1.8. *Let* (X, d) *and* (Y, δ) *be two metric space. We equip* $\mathsf{X} \times \mathsf{Y}$ *with the metric* $d + \delta$. *Let* X, X_n, $n \geq 1$ *be random variables valued in* $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* Y_n, $n \geq 1$ *be random variables valued in* $(\mathsf{Y}, \mathcal{B}(\mathsf{X}))$ *and defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *The following assertions hold.*

*(*i*) If* $X_n \xrightarrow{\text{a.s.}} X$, *then* $X_n \xrightarrow{P} X$.

*(*ii*) If* $X_n \xrightarrow{P} X$, *then* $X_n \Longrightarrow X$.

*(*iii*) For all* $c \in \mathsf{X}$, $X_n \xrightarrow{P} c$ *if and only if* $X_n \Longrightarrow c$,

*(*iv*) Suppose that the spaces* (X, d) *and* (Y, δ) *coincide. If* $X_n \Longrightarrow X$ *and* $d(X_n, Y_n) \xrightarrow{P} 0$, *then* $Y_n \Longrightarrow X$.

*(*v*) For all* $c \in \mathsf{X}$, *if* $X_n \Longrightarrow X$ *and* $Y_n \xrightarrow{P} c$, *then* $(X_n, Y_n) \Longrightarrow (X, c)$.

*(*vi*) If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then* $(X_n, Y_n) \xrightarrow{P} (X, Y)$.

The following classical lemma can be useful.

Lemma C.1.9. *Let $(Z_{n,m})_{n,m\geq 1}$ be an array of random variables in X. Suppose that for all $m \geq 1$, $Z_{n,m}$ converges weakly to Z_m as $n \to \infty$ and that Z_m converges weakly to Z as $m \to \infty$. Let now $(X_n)_{n\geq 1}$ be random variables in X such that, for all $\epsilon > 0$,*

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}(d(X_n, Z_{m,n}) > \epsilon) = 0 \, .$$

Then X_n converges weakly to Z as $n \to \infty$.

Proof. Let $f \in \mathrm{Lip}_b(\mathsf{X})$ so that $|f(x) - f(y)| \leq K\, d(x, y)$ and $|f(x)| \leq C$ for all $x, y \in \mathsf{X}$. Then we write

$$\begin{aligned}\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right] = \mathbb{E}\left[f(X_n) - f(Z_{m,n})\right] \\ + \left[\mathbb{E}\left[f(Z_{m,n})\right] - \mathbb{E}\left[f(Z_m)\right]\right] + \left[\mathbb{E}\left[f(Z_m)\right] - \mathbb{E}\left[f(Z)\right]\right] . \quad \text{(C.3)}\end{aligned}$$

Then, for all $\epsilon > 0$, since $|f(X_n) - f(Z_{m,n})| \leq K\,\epsilon$ if $d(X_n, Z_{m,n}) \leq \epsilon$ and $|f(X_n) - f(Z_{m,n})| \leq C$ otherwise, we have,

$$\mathbb{E}\left[|f(X_n) - f(Z_{m,n})|\right] \leq K\,\epsilon + C\,\mathbb{P}(d(X_n, Z_{m,n}) > \epsilon)$$

By Theorem C.1.4 and using the assumptions of the lemma, we get that, for some large enough m,

$$\limsup_{n\to\infty} \left|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right]\right| \leq (K\,\epsilon + C)\epsilon + 0 + \epsilon \, .$$

Hence, since $\epsilon > 0$ can be taken arbitrarily small, $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(Z)\right]$ and we conclude with Theorem C.1.4. □

C.2 Some topology results

An important fact about the weak convergence on $\mathbb{M}_1(\mathsf{X})$ is that it is metrizable, provided that X is separable. This is shown in the two following results.

Let us denote by $\mathcal{S}$ the class of closed sets of X and, for $A \subset \mathsf{X}$ and $\alpha > 0$, $A^\alpha = \{x \in \mathsf{X},\, d(x, A) < \alpha\}$. A^α is an open set and $A^\alpha \downarrow \overline{A}$ if $\alpha \downarrow 0$. We set, for $\lambda, \mu \in \mathbb{M}_1(\mathsf{X})$,

$$\boldsymbol{\rho}(\lambda, \mu) = \inf\left\{\alpha > 0 \,:\, \lambda(F) \leq \mu(F^\alpha) + \alpha \text{ for all } F \in \mathcal{S}\right\} . \quad \text{(C.4)}$$

The following result shows that $\boldsymbol{\rho}$ is indeed a metric, which is not completely obvious from (C.4).

Lemma C.2.1. *$\boldsymbol{\rho}$ defined in (C.4) is a metric on $\mathbb{M}_1(\mathsf{X})$.*

The following result indicates that the metric $\boldsymbol{\rho}$ defines the topology of the weak convergence whenever (X, d) is separable.

Proposition C.2.2. *Assume that (X, d) is separable. Let $(\mu_n)_{n\in\mathbb{N}} \subset \mathbb{M}_1(\mathsf{X})$ and $\mu \in \mathbb{M}_1(\mathsf{X})$. Then $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ iff $\boldsymbol{\rho}(\mu_n, \mu) \to 0$. Moreover $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is separable.*

In the following, we assume that (X, d) is separable, so that, by Proposition C.2.2, $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is a separable metric space associated to the weak convergence. As a consequence, a subset $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is compact if it is sequentially compact.

The relative compactness of a subset of $\mathbb{M}_1(\mathsf{X})$ can be related to its *tightness*, that is, coarsely speaking, the property of all the measures of this subset to be almost supported on the same compact subset of X.

Definition C.2.1. *Let Γ be a subset of $\mathbb{M}_1(\mathsf{X})$.*

(i) We say that Γ is tight if for all $\epsilon > 0$, there exists a compact set $K \subset \mathsf{X}$ such that $\mu(K) \geq 1 - \epsilon$ for all $\mu \in \Gamma$.

(ii) We say that Γ is relatively compact if every sequence of elements in Γ contains a weakly convergent subsequence, or, equivalently if $\overline{\Gamma}$ is compact.

The following result is often referred to as the *Prokhorov theorem*.

Theorem C.2.3. *Let (X, d) be separable. Then if $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is tight, it is relatively compact.*

This theorem has the following converse result in the case where (X, d) is complete.

Theorem C.2.4. *Let (X, d) be separable and complete. If $\Gamma \subset \mathbb{M}_1(\mathsf{X})$ is relatively compact, then it is tight.*

Since singletons are compact, a direct but important consequence of this theorem is that any $\{\mu\} \subset \mathbb{M}_1(\mathsf{X})$ is tight.

Let us conclude this section with a last topological result.

Theorem C.2.5. *Let (X, d) be separable and complete. Then $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ is separable and complete.*

Index

Bibliography

M.A. Al-Osh and A.A. Alzaid. First order integer-valued autoregressive (inar (1)) process. *J. Time Series Anal*, 8:261–275, 1987.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, NY, third edition, 2003.

P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999a. A Wiley-Interscience Publication.

Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 1999b. ISBN 0-471-19745-9.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 2nd edition, 1991.

P. E. Caines. *Linear Stochastic Systems*. Wiley, 1988.

Ju. A. Davydov. The invariance principle for stationary processes. *Teor. Verojatnost. i Primenen.*, 15:498–509, 1970. ISSN 0040-361x.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

C. W. J. Granger and Roselyne Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1980.tb00297.x. URL http://dx.doi.org/10.1111/j.1467-9892.1980.tb00297.x.

C.W.J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121 – 130, 1981. ISSN 0304-4076. doi: http://dx.doi.org/10.1016/0304-4076(81)90079-8. URL http://www.sciencedirect.com/science/article/pii/0304407681900798.

E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. John Wiley & Sons, 1988.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*. Springer Series in Statistics. Springer, New York, 2012. ISBN 978-1-4614-3654-6. doi: 10.1007/978-1-4614-3655-3. URL http://dx.doi.org/10.1007/978-1-4614-3655-3.

Jean Jacod and Philip Protter. *Probability essentials*. Universitext. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43871-8.

R. E. Kalman and R. Bucy. New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME, Series D*, 83(3):95–108, 1961.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D):35–45, 1960.

J. F. C. Kingman. Subadditive ergodic theory. *Ann. Probability*, 1:883–909, 1973. With discussion by D. L. Burkholder, Daryl Daley, H. Kesten, P. Ney, Frank Spitzer and J. M. Hammersley, and a reply by the author.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Cambridge University Press, 2005.

Magda Peligrad and Sergey Utev. Invariance principle for stochastic processes with short memory. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 18–32. Inst. Math. Statist., Beachwood, OH, 2006. doi: 10.1214/074921706000000734. URL http://dx.doi.org/10.1214/074921706000000734.

P.M. Robinson. Log-periodogram regression of time series with long range dependence. *Ann. Statist.*, 23:1043–1072, 1995.

H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988. ISBN 0-02-404151-3.

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications*. New York: Springer, 3rd edition, 2011.

R software. The R project for statistical computing. http://www.r-project.org/.

R.S. Tsay. *Analysis of Financial Time Series*, volume 543. Wiley-Interscience, 2005.

Nicholas Young. *An introduction to Hilbert space*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1988. ISBN 0-521-33071-8; 0-521-33717-8.