Linear time series
TSIA202

François Roueff

August 19, 2019

Contents

Foreword

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.

The main goal of these lecture notes is to provide a solid introduction to the basic principles of stochastic modeling, statistical inference and forecasting methods for time series. Essential references for students interested in these topics are [3] and [9]. In these notes, we will mainly consider linear models. We will start by setting the general framework of stochastic modelling in Chapter 1. We will focus on second order properties in Chapter 2 and linear models in Chapter 3, with a detailed description of the ARMA model. In Chapter 4, we will study the most widespread statistical approaches for linear forecasting. Numerical algorithms for forecasting will be derived in this context. Finally Chapter 6 contains the main results for second order statistical inference of linear models. It includes a series of results about the asymptotic behavior of standard estimators in time series. These results are more involved and can be omitted on a first reading, especially if the standard results on the convergence of random variables recalled in Appendix A are not well known to the reader. In the latter case, we recommend to focus on Section 6.2.

Notation and conventions

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

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

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 $X_n \overset{\text{fidi}}{\Longrightarrow} X$.

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/*.

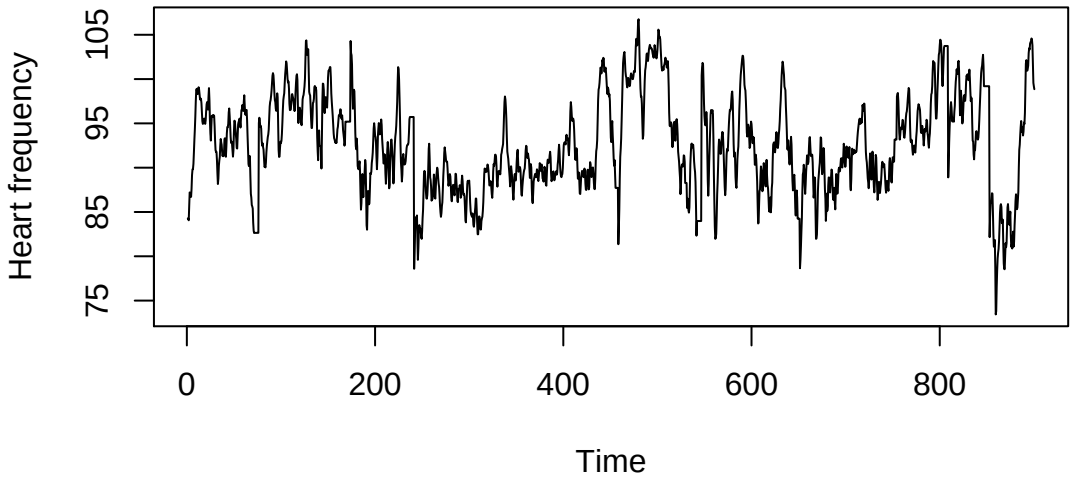

Figure 1.1: *Heartbeats: time evolution of the heart rate.*

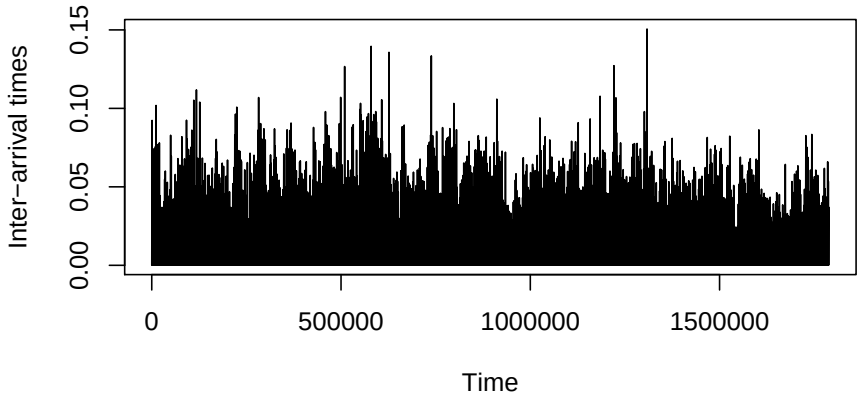

Figure 1.2: *Internet traffic trace : inter-arrival times of TCP packets.*

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).*

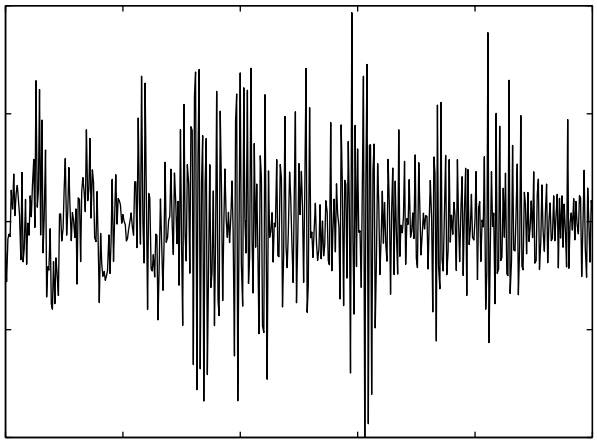

Figure 1.3: *A record of the unvoiced fricative phoneme* sh*.*

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

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.*

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 $T = \mathbb{Z}$.

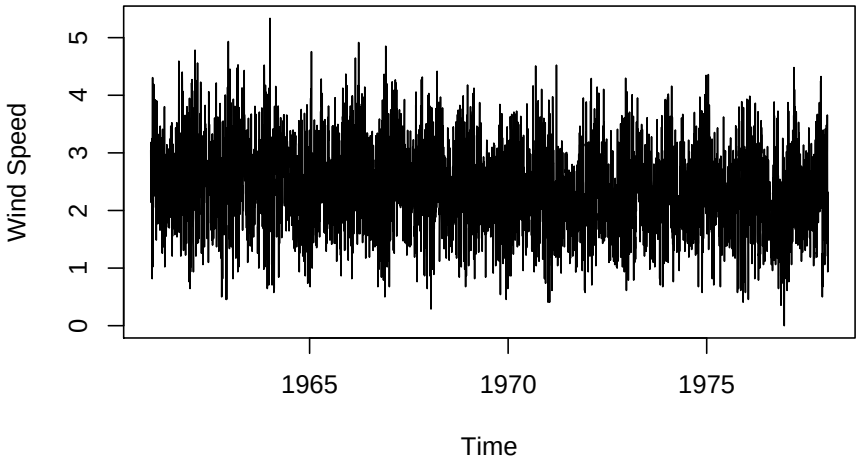

Figure 1.4: *Daily record of the wind speed at Kilkenny (Ireland).*

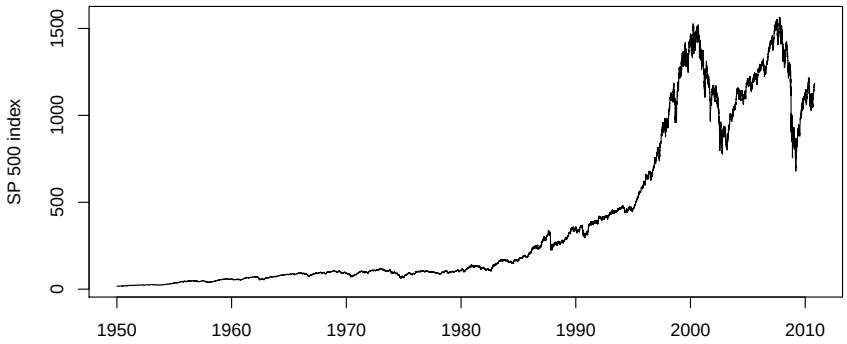

Figure 1.5: SP-500 stock index time series

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.*

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

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

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 (see [6, Corollaire 6.1]).

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 $(\prod_{t\in T} \mathsf{X}_t, \bigotimes_{t\in T} \mathcal{X})$.

If T is infinite, this definition 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

$$\mathcal{X}^{\otimes T} = \sigma\left\{\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I} : I \in \mathcal{I}, \forall t \in I,\ A_t \in \mathcal{X}\right\} , \tag{1.1}$$

where $\mathcal{I}$ denotes the set of finite subsets of T.

Let $X = (X_t)_{t\in T}$ be random process $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$ and $I \in \mathcal{I}$. Let $\mathbb{P}_I$ 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

$$\mathbb{P}_I\left(\prod_{t\in I} A_t\right) = \mathbb{P}\left(X_t \in A_t,\ t \in I\right) , \tag{1.2}$$

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

Definition 1.2.3. *We call* finite dimensional distributions *or* fidi distributions *of the process X the collection of probability measures $(\mathbb{P}_I)_{I\in\mathcal{I}}$.*

The probability measure $\mathbb{P}_I$ is sufficient to compute the probability of any event of the form $\mathbb{P}(\cap_{t\in I}\{X_t \in A_t\})$ where $\{A_t, t \in I\} \subset \mathcal{X}$, or, equivalently, to compute the expectation $\mathbb{E}\left[\prod_{t\in I} f_t(X_t)\right]$ where for all $t \in I$, f_t is a non-negative measurable function.

Let Π_I denote the canonical projection of X^T on X^I,

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

If $I = \{s\}$ with $s \in T$, we will denote

$$\Pi_s(x) = \Pi_{\{s\}}(x) = x_s \quad \text{for all} \quad x = (x_t)_{t\in T} \in \mathsf{X}^T \ . \tag{1.4}$$

The fidi distributions of a given process from a single probability measure on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$, called the *law* of the process in the sense of fidi distributions, and defined as follows.

Definition 1.2.4 (Law of a random process on $\mathcal{X}^{\otimes T}$)**.** *Let $X = (X_t)_{t\in T}$ be a random process defined on $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{X}, \mathcal{X})$. The* law in the sense of fidi distributions *is the* image measure $\mathbb{P}^X$*, that is, the unique probability measure defined on $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$ that satisfies $\mathbb{P}^X \circ \Pi_I^{-1} = \mathbb{P}_I$ for all $I \in \mathcal{I}$, i.e.*

$$\mathbb{P}^X\left(\prod_{t\in I} A_t \times \mathsf{X}^{T\setminus I}\right) = \mathbb{P}\left(X_t \in A_t,\, t \in I\right)$$

for all $(A_t)_{t\in I} \in \mathcal{X}^I$.

One can see $\mathbb{P}^X$ as the law of a random variable valued in $(\mathsf{X}^T, \mathcal{X}^{\otimes T})$.

We will always admit that the random process $X = (X_t)_{t\in T}$ can indeed be constructed for a given "well chosen" distribution. One can in fact provide a simple criterion on the collection $(\mathbb{P}_I)_{I\in\mathcal{I}}$ to ensure that such a construction is valid but it is not the object of this course to focus on this matter. We will satisfy ourselves here with the following important example (whose existence is admitted). All other examples will be constructed from them.

Example 1.2.1 (Independent processes)**.** *Let $(\nu_t)_{t\in T}$ be a collection of probability measures on $(\mathsf{X}, \mathcal{X})$. We say that $X = (X_t)_{t\in T}$ is an* independent process *with marginals $(\nu_t)_{t\in T}$ if $(X_t)_{t\in T}$ is a collection of independent random variables and $X_t \sim \nu_t$ for all $t \in T$. In that case, for all $I \in \mathcal{I}$, we have*

$$\nu_I = \bigotimes_{t\in I} \nu_t \ , \tag{1.5}$$

where $\otimes$ denotes the tensor product of measures, that is,

$$\nu_I\left(\prod_{t\in T} A_t\right) = \prod_{t\in T} \nu_t(A_t) \ .$$

We say that $X = (X_t)_{t\in T}$ *is an* i.i.d. (independent and identically distributed) process *if moreover* ν_t *does not depend on* t.

1.2.3 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 [6, Chapter 16].

Definition 1.2.5 (Gaussian variable)**.** *The real valued random variable* X *is Gaussian if its characteristic function satisfies :*

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

where $\mu \in \mathbb{R}$ *and* $\sigma \in \mathbb{R}_+$.

One can show that $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. If $\sigma \neq 0$, then X admits a probability density function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{1.6}$$

If $\sigma = 0$, then $X = \mu$ a.s. This definition can be extended to random vectors as follows.

Definition 1.2.6 (Gaussian vector)**.** *A random vector* $[X_1, \ldots, X_n]^T$ *valued in* $\mathbb{R}^n$ *is 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

$$\phi_X(u) = \mathbb{E}\left[\exp(\mathrm{i}u^T X)\right] = \mathbb{E}\left[\exp(\mathrm{i}Y)\right] = \exp\left(\mathrm{i}u^T\mu - \frac{1}{2}u^T\Gamma u\right) \tag{1.7}$$

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 1.2.1. *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 symmetric matrix Γ, the distribution $X \sim \mathcal{N}(\mu, \Gamma)$ is well defined.

Proof. The first part of the result follows directly from (1.7). It also yields the following lemma.

Lemma 1.2.2. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ being a $n \times n$ non-negative 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 1.2.1, just observe that all non-negative 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 [6, Corollaire 16.1]).

Proposition 1.2.3. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ a $n \times n$ non-negative symmetric matrix. Then X has independent components if and only if Γ is diagonal.*

Using the same path as in the proof of Proposition 1.2.1, 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 [6, Corollaire 16.2]).

Proposition 1.2.4. *Let $X \sim \mathcal{N}(\mu, \Gamma)$ with $\mu \in \mathbb{R}^n$ and Γ a $n \times n$ non-negative 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.

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

Definition 1.2.7 (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 (1.7), for all finite set of indices $I = \{t_1, t_2, \cdots, t_n\}$, the finite distribution ν_I is the Gaussian probability on $\mathbb{R}^n$

$$\nu_I \stackrel{\text{def}}{=} \mathcal{N}(\mu_I, \Gamma_I) \tag{1.8}$$

where $\mu_I = [\mu(t_1), \ldots, \mu(t_n)]^T$, $\Gamma_I(m,k) = \gamma(t_m, t_k)$, and where we used 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, the matrix Γ_I with entries, with $1 \le m, k \le n$, is a covariance matrix of a random vector of dimension n. It is therefore nonnegative 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, we admit that there exists a Gaussian process having this functions as mean and covariance functions as stated hereafter.

Theorem 1.2.5. *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 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 1.2.8 (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 1.2.5. 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.*

1.3 Strict stationarity of a random process in discrete time

1.3.1 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 1.3.1 (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 1.3.2 (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 $I \in \mathcal{I}$,

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

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 $\mathbb{P}_I = \mathbb{P}_{I+\tau}$, where $I + \tau = \{t + \tau, t \in I\}$.

Example 1.3.1 (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 $t_1, \ldots, t_n$. Observe 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 $t \in T$.*

1.3.2 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 1.3.2 (Moving transformation of an i.i.d. process)**.** *Let Z be an i.i.d. process (see Example 1.3.1). 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.*

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 1.3.3. *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 defined in (1.4). 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 1.3.2, 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 1.3.3 (Shift)**.** *A very basic linear filtering is obtained with $\phi = \mathrm{S}$ where S is the shift operator of Definition 1.3.1. In this case $Y_t = X_{t+1}$ for all $t \in \mathbb{Z}$.*

Example 1.3.4 (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 1.3.5 (Differencing operator)**.** *A particular case is the* differencing operator $I - \mathrm{S}^{-1}$ *where* I *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 $Y = (I - \mathrm{S}^{-1})^k X$ *is given by*

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

Example 1.3.6 (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 1.3.4. *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 1.3.1. *A* shift invariant ϕ*-filter is entirely determined by its composition with the canonical projection* Π_0 *defined in (1.4). 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.*

[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.

1.4 Exercises

Exercise 1.1. 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 1.2. 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 1.3. 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 1.2.1].

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 1.4. 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 $|\rho| < 1$.
4. $Y_t = \varepsilon_t \varepsilon_{t-1}$.
5. $Y_t = (-1)^t \varepsilon_t$, $Z_t = \varepsilon_t + Y_t$.

Chapter 2

Weakly stationary processes

In this chapter, we focus on second order properties of 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).

2.1 L^2 processes

We will often denote the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ of $\mathbb{C}^d$-valued random variables with finite variance,

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) = \left\{ X \text{is a } d\text{-dimensional r.v. on } (\Omega, \mathcal{F}, \mathbb{P}) \text{ s.t. } \mathbb{E}\left[|X|^2\right] < \infty \right\} ,$$

simply as L^2. (Note that d does not appear in the notation, but we will essentially consider the case $d = 1$).

Definition 2.1.1 (L^2 Processes)**.** *The process $\mathbf{X} = (X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{C}^d$ is an L^2 process if $\mathbf{X}_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ for all $t \in T$.*

The *mean function* defined on T by $\boldsymbol{\mu}(t) = \mathbb{E}\left[\mathbf{X}_t\right]$ takes its values in $\mathbb{C}^d$ and the *covariance function* is defined on $T \times T$ by

$$\Gamma(s,t) = \mathrm{Cov}\left(\mathbf{X}_s, \mathbf{X}_t\right) = \mathbb{E}\left[(\mathbf{X}_s - \boldsymbol{\mu}(s))(\mathbf{X}_t - \boldsymbol{\mu}(t))^H\right] ,$$

which takes its values in $d \times d$ matrices. We will sometimes use the notation $\boldsymbol{\mu}_{\mathbf{X}}$ and $\Gamma_{\mathbf{X}}$, the subscript $\mathbf{X}$ indicating the process used in these definitions. For all $s \in T$, $\Gamma(s,s)$ is a covariance matrix and is thus nonnegative definite hermitian. More generally, the following properties hold.

Proposition 2.1.1. *Let Γ be the covariance function of a L^2 process $\mathbf{X} = (\mathbf{X}_t)_{t\in T}$ with values in $\mathbb{C}^d$. The following properties hold.*

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

$$\Gamma(s,t) = \Gamma(t,s)^H \tag{2.1}$$

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

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

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

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

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

which implies (2.2).

The converse assertion follows from Exercise 1.3. □

In the scalar case ($d = 1$), we will also use the notation $\gamma(s,t)$.

2.2 Weakly stationary processes

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 $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathbb{Z}}$ be a strictly stationary L^2 process with values in $\mathbb{C}^d$. 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 2.2.1 (Weakly stationary processes)**.** *Let $\boldsymbol{\mu} \in \mathbb{C}^d$ and $\Gamma : \mathbb{Z} \to \mathbb{C}^{d\times d}$. A process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}^d$ is said* weakly stationary *with mean $\boldsymbol{\mu}$ and autocovariance function Γ if all the following assertions hold:*

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

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

(iii) for all $(s,t) \in \mathbb{Z}\times\mathbb{Z}$, $\mathrm{Cov}\,(\mathbf{X}_s, \mathbf{X}_t) = \Gamma(s-t)$.

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 1.2.3, see Proposition 1.2.1.

Observe that a process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}^d$ is weakly stationary with mean $\boldsymbol{\mu}$ and autocovariance function Γ if and only if for all $\lambda \in \mathbb{C}^d$, the process $(\lambda^H \mathbf{X}_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}$ is weakly stationary with mean $\lambda^H\boldsymbol{\mu}$ and autocovariance function $\lambda^H\Gamma\lambda$. The study of weakly stationary processes can thus be done in the case $d = 1$ without a great loss of generality.

2.2.1 Properties of the autocovariance function

The properties of Proposition 2.1.1 imply the following ones in the case of a weakly stationary process.

Proposition 2.2.1. *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 2.2.2 (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,

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

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 2.2.3 (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 2.2.1 (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 2.2.2 (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 2.2.3 (Random walk)**.** *Let (S_t) be a random process defined on $t \in \mathbb{N}$ by $S_t = X_0 + X_1 + \cdots + X_t$, where (X_t) is a strong white noise. 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) is not weakly stationary.

Example 2.2.4 (Continued from Example 2.2.1)**.** *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 2.2.1 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 2.2.5 (Invariance of the autocovariance function under time reversion (continued from Example 1.3.6)**.** *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 1.3.6. Then (Y_t) is weakly stationary with same mean as X and autocovariance function $\gamma_Y = \overline{\gamma_X}$.*

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

2.2.2 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 2.2.6 (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, recalling that L^2 is endowed with a scalar product, we have the projection formula

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

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.

2.3 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 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.

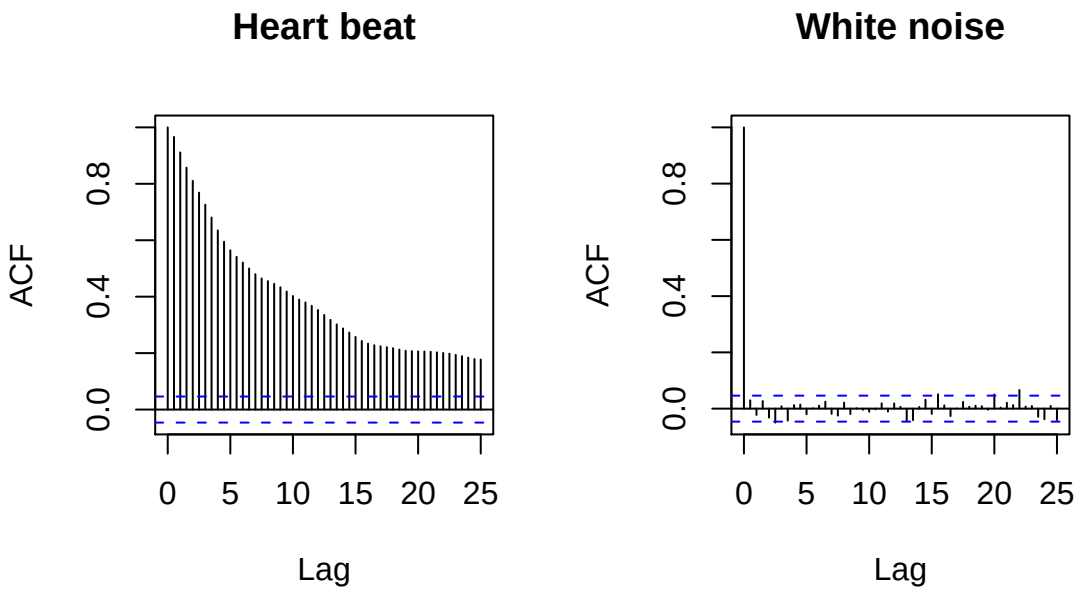

Figure 2.1: *Left : empirical autocorrelation* $\hat{\rho}_n(h)$ *of heartbeat data for* $h = 0, \ldots, 100$. *Right : the same from a simulated white noise sample with same length.*

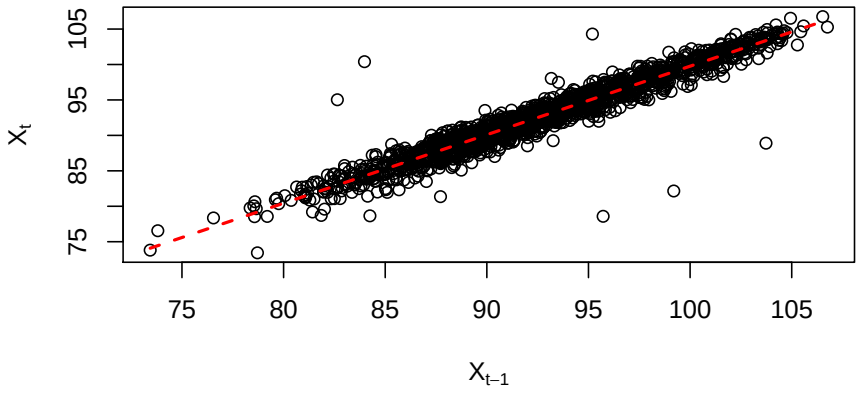

Figure 2.2: *Each point is a couple* (X_{t-1}, X_t), *where* $X_1, \ldots, X_n$ *is the heartbeat data sample. The dashed line is the best approximation of* X_t *as a linear function of* X_{t-1}.

Theorem 2.3.1 (Herglotz)**.** *A sequence $(\gamma(h))_{h\in\mathbb{Z}}$ is a nonnegative definite hermitian sequence in the sense of Proposition 2.2.1 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 2.3.1. *By Proposition 2.2.1, Theorem 2.3.1 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

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

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

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

We can multiply the sequence (ν_n) by a constant to obtain a sequence of probability measures. Thus Theorem A.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 $h \in \mathbb{Z}$: $\int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}h\lambda} \mu(\mathrm{d}\lambda)$. The uniform convergence of Fourier series in the Cesaro sense (see [11]) tells us that any continuous (2π)-periodic function g can be approximated uniformly by

$$\frac{1}{n} \sum_{k=0}^{n-1} g_k \quad \text{with} \quad g_k = \sum_{j=-k}^{k} \left(\frac{1}{2\pi} \int_{\mathbb{T}} g(\lambda) \mathrm{e}^{-\mathrm{i}j\lambda} \mathrm{d}\lambda \right) \mathrm{e}^{\mathrm{i}j\lambda} \ .$$

We thus obtain that $\int_{\mathbb{T}} g(\lambda)\nu(\mathrm{d}\lambda) = \int_{\mathbb{T}} g(\lambda)\mu(\mathrm{d}\lambda)$. Since this true for all such g's, this implies $\nu = \mu$. □

Corollary 2.3.2 (The ℓ^1 case)**.** *Let $(\gamma(h))_{h\in\mathbb{Z}} \in \ell^1(\mathbb{Z})$. Then it is a nonnegative definite hermitian sequence in the sense of Proposition 2.2.1 if and only if*

$$f(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma(h) \mathrm{e}^{-\mathrm{i}h\lambda} \geq 0 \ ,$$

for all $\lambda \in \mathbb{T}$.

Proof. Left as an exercise. □

Exercise 2.1. Prove Corollary 2.3.2 (apply Theorem 2.3.1 and the definition of f).

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

Example 2.3.1 (MA(1), Continued from Example 2.2.4)**.** *Consider Example 2.2.4. 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.

Example 2.3.2 (Spectral density function of a white noise)**.** *Recall the definition of a white noise, Definition 2.2.3. 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.

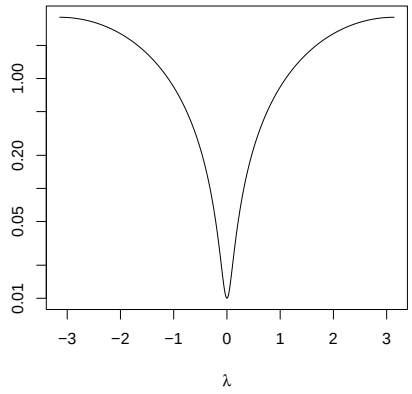

Figure 2.3: *Spectral density function (in logarithmic scale) of an MA(1) process, as given by (2.4) with* $\sigma = 1$ *and* $\theta = -0.9$.

Example 2.3.3 (Spectral measure of an harmonic process, continued from Example 2.2.2)**.** *The autocovariance function of* X *is given by (see Example 2.2.2)*

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

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 2.3.1 (Linearly predictable processes)**.** *A weakly stationary process* X *is called* 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 2.3.3. *Let γ be the autocovariance function of a weakly stationary process X. If $\gamma(0) \neq 0$ and $\gamma(t) \to 0$ as $t \to \infty$ then X is not linearly predictable.*

2.4 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

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

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 $\mathcal{H}_\infty^X$ as follows

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

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

Definition 2.4.1 (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*

$$\epsilon_t = X_t - \mathrm{proj}\left(X_t | \,\mathcal{H}_{t-1}^X\right)\,. \tag{2.13}$$

By the orthogonal principle of projections in L^2, 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

$$\sigma^2 = \|\epsilon_t\|^2 = \mathbb{E}\left[|\epsilon_t|^2\right] \tag{2.14}$$

does not depend on t.

Example 2.4.1 (Innovation process of a white noise)**.** *The innovation process of a white noise $X \sim \mathrm{WN}(0, \sigma^2)$ is $\epsilon = X$.*

Example 2.4.2 (Innovation process of a MA(1), continued from Example 2.2.1)**.** *Consider the process X defined in Example 2.2.1. 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 (2.13) 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 $p \geq 0$, the finite dimensional space

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

and observe that $(\mathcal{H}_{t,p}^X)_p$ is an increasing sequence of linear space whose union has closure $\mathcal{H}_t^X$. In this case we have, for any L^2 variable Y,

$$\lim_{p\to\infty} \mathrm{proj}\left(Y|\,\mathcal{H}_{t,p}^X\right) = \mathrm{proj}\left(Y|\,\mathcal{H}_t^X\right) \ , \tag{2.15}$$

where the limit holds in the L^2 sense.

Definition 2.4.2 (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* $\mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,p}^X\right)$ *and the* partial innovation process of order p *the process $\epsilon_p^+ = \left(\epsilon_{t,p}^+\right)_{t\in\mathbb{Z}}$ defined by*

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

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

$$\mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1,p}^X\right) = \sum_{k=1}^{p} \phi_{k,p}^+ X_{t-k} \ . \tag{2.16}$$

Observe that, by the orthogonality principle, (2.16) is equivalent to

$$\Gamma_p^+ \phi_p^+ = \gamma_p^+ \ , \tag{2.17}$$

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

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

Observing that Equation (2.17) 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

$$\sigma_p^2 = \| X_t - \mathrm{proj}\,(X_t | \mathcal{H}_{t-1,p}) \|^2 = \mathbb{E}\left[| X_t - \mathrm{proj}\,(X_t | \mathcal{H}_{t-1,p}) |^2 \right] . \quad (2.18)$$

By (2.16) and the usual orthogonality condition of the projection, we have

$$\begin{aligned} \sigma_p^2 &= \langle X_t, X_t - \mathrm{proj}\,(X_t | \mathcal{H}_{t-1,p}) \rangle \\ &= \gamma(0) - \sum_{k=1}^{p} \overline{\phi^+_{k,p}} \gamma(k) \\ &= \gamma(0) - (\phi^+_p)^H \gamma^+_p . \end{aligned} \quad (2.19)$$

Equations (2.17) and (2.19) are called *Yule-Walker equations*. An important consequence of these equations is that σ_p^2 does not depend on t, and since (2.15) implies

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

we obtain that, as claimed above, the variance of the innovation defined in (2.14) is also independent of t. So we can state the following result.

Corollary 2.4.1. *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 2.4.3 (Innovations of the harmonic process (continued from Example 2.2.2))**.** *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 $\sigma_2^2 = \| X_t - \mathrm{proj}\left(X_t | \mathcal{H}^X_{t-1,2}\right) \|^2 = 0$ and thus

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

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 2.4.3 (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

$$\mathcal{H}^X_{-\infty} = \bigcap_{t\in\mathbb{Z}} \mathcal{H}^X_t \ .$$

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

Example 2.4.4 (Constant process)**.** *A very simple example of deterministic process is obtained by taking $\lambda_0 = 0$ in Example 2.4.3. 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 $\mathcal{H}^X_{-\infty} = \{0\}$. 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 $t \in \mathbb{Z}$

$$\mathcal{H}^Z_t \subseteq \mathcal{H}^X_t \overset{\perp}{\oplus} \mathcal{H}^Y_t \ .$$

This implies that

$$\mathcal{H}^Z_{-\infty} \subseteq \mathcal{H}^X_{-\infty} \overset{\perp}{\oplus} \mathcal{H}^Y_{-\infty} \ . \tag{2.20}$$

Also, by the orthogonality principle of the projection, 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 2.4.5 (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 (2.20), $\mathcal{H}^Z_{-\infty} \subseteq \mathrm{Span}\,(Y_0)$. 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 2.4.5 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,

$$\psi_s = \frac{\langle X_t, \epsilon_{t-s} \rangle}{\sigma^2} \ , \tag{2.21}$$

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 $t \in \mathbb{Z}$,

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

We can now state the Wold decomposition, whose proof is admitted here.

Theorem 2.4.2 (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 (2.21). Defined the L^2 centered process V by the following equation:

$$X_t = U_t + V_t, \quad \textit{for all } t \in \mathbb{Z}. \tag{2.22}$$

Then the following assertions hold.

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

2.5 Exercises

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 1.4, 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 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 f and denote by $\hat{X}$ its spectral representation field, 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 2.3.3 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 2.4.5]

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 2.4.2.

1. Show that
$$\mathcal{H}^X_{-\infty} \overset{\perp}{\oplus} \mathcal{H}^\epsilon_t = \mathcal{H}^X_t \ .$$

2. Deduce that $U_t = \mathrm{proj}\,(X_t|\,\mathcal{H}^\epsilon_t)$, $V_t = \mathrm{proj}\left(X_t|\,\mathcal{H}^X_{-\infty}\right)$ 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 2.4.2.

Chapter 3

Linear 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 representations but this is out of the scope of these lecture notes.

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 1.3.1. 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 2.3.1) 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*

$$\Gamma_1 = \{z \in \mathbb{C} : |z| = 1\} .$$

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 $\Delta_1 = \{z \in \mathbb{C} : |z| \le 1\}$. 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 $z \in \Gamma_1$ (the unit circle).

Proof. The only fact to show is that (3.11) on $z \in \Gamma_1$ 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 1.3.6.

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

$$\mathrm{F}_{\phi} \circ \mathrm{F}_{\tilde{\phi}} = I \; . \tag{3.16}$$

Here I 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 2.2.1, 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 $\phi_1, \ldots, \phi_p \in \mathbb{C}$. 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 Γ_1. Moreover, in this case, we have $X = \mathrm{F}_\psi(Z)$, where $\psi \in \ell^1$ is uniquely defined by

$$\sum_{t\in\mathbb{Z}} \psi_t z^t = \frac{1}{\Phi(z)} \quad \text{on} \quad z \in \Gamma_1 \ .$$

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

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

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 $\phi_1, \ldots, \phi_p$ and MA coefficients $\theta_1, \ldots, \theta_q$ 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 $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q \in \mathbb{C}$. 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 Γ_1. Moreover, in this case, we have $X = \mathrm{F}_\psi(Z)$, where $\psi \in \ell^1$ is uniquely defined by*

$$\sum_{t\in\mathbb{Z}} \psi_t z^t = \frac{\Theta}{\Phi}(z) \quad \textit{on} \quad z \in \Gamma_1 \,. \tag{3.28}$$

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

$$\sum_{t\in\mathbb{Z}} \xi_t z^t = \frac{1}{\Phi(z)} \quad \text{on} \quad z \in \Gamma_1 \,.$$

(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 $\epsilon > 0$,

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

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 $\tilde{\phi}_1, \dots, \tilde{\phi}_p$ and MA coefficients $\tilde{\theta}_1, \dots, \tilde{\theta}_q$ 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 the orthogonality principle of projection, 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 $\mathrm{WN}(0,\sigma^2)$, see Section 2.4. Since $\gamma(h) = 0$ for all $|h| > q$, we have $X_t \perp \mathcal{H}_{t-q-1}^X$ for all t. Observing that

$$\mathcal{H}_{t-1}^X = \mathcal{H}_{t-q-1}^X \overset{\perp}{\oplus} \mathrm{Span}\left(\epsilon_{t-q}, \ldots, \epsilon_{t-1}\right) \ ,$$

by the orthogonality principle of projection, we obtain that

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

and thus, either $\sigma^2 = 0$ and $X_t = 0$ a.s. (a very trivial MA process) or X is regular and we have

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1}^X\right) = \sum_{k=1}^{q} \frac{\langle X_t, \epsilon_{t-k}\rangle}{\sigma^2}\, \epsilon_{t-k} \ .$$

where the coefficient in front of each ϵ_{t-k} does not depend on t, but only on k (see (2.21). Let us presently denote it by θ_k. Since $X_t = \mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1}^X\right) + \epsilon_t$, 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 $t \in \mathbb{Z}$,

$$\mathrm{proj}\left(X_t \middle| \mathcal{H}_{t-1,m}^X\right) = \sum_{k=1}^{p} \phi_k X_{t-k} \ .$$

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 $t \in \mathbb{Z}$,

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

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 $t \in \mathbb{Z}$,

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,m}\right) \in \mathcal{H}^X_{t-1,m-1} \ ,$$

which implies that

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,m}\right) = \operatorname{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,m-1}\right) \ ,$$

and, iterating in m,

$$\operatorname{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,m}\right) = \operatorname{proj}\left(X_t \middle| \mathcal{H}^X_{t-1,p}\right) \ .$$

Letting $m \to \infty$, by (2.15), we get that

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

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 2.4.1) 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

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \ , \qquad (3.31)$$

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 1: 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 1 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 1 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 2: 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 1 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 Beyond absolutely summable coefficients

Let us conclude with an example where, given a weakly stationary process X and a random variable in $\mathcal{H}_\infty^X$, one defines a linear filter without relying on a sequence of absolutely summable coefficients.

Example 3.7.1 (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 $n \to \infty$,*

$$\sum_{s\in\mathbb{Z}} \alpha_{n,s} X_{-s} \to Y_0 \quad \text{in } L^2.$$

It follows that, by weak stationarity and using the Cauchy criterion, for all $t \in \mathbb{Z}$,

$$\sum_{s\in\mathbb{Z}} \alpha_{n,s} X_{t-s} \to Y_t \quad \text{in } L^2,$$

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 instance of the previous case is obtained when X is a white noise.

Example 3.7.2 (The white noise case)**.** *We consider Example 3.7.1 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

$$Y_t = \sum_{s\in\mathbb{Z}} \alpha_s X_{t-s} \quad \text{in } L^2, \tag{3.37}$$

and

$$\gamma'(\tau) = \sum_{s\in\mathbb{Z}} \sum_{t\in\mathbb{Z}} \alpha_s\alpha_t\gamma(\tau - t + s) \, .$$

3.8 Exercises

Exercise 3.1. Suppose that

$$Y_t = \beta t + S_t + X_t, \quad t \in \mathbb{Z}\,,$$

where $\beta \in \mathbb{R}$, $(S_t)_{t\in\mathbb{Z}}$ is a 4-periodic weakly stationary process and $(X_t)_{t\in\mathbb{Z}}$ is a weakly stationary process such that (X_t) and (S_t) are uncorrelated.

1. Is (Y_t) weakly stationary ?
2. Which property is satisfied by the covariance function of (S_t) ? Define $(\bar{S}_t)$ as the process obtained by applying the operator $1+\mathrm{B}+\mathrm{B}^2+\mathrm{B}^3$ to (S_t), where B denotes the shift operator. What can be said about $(\bar{S}_t)$?
3. Consider now (Z_t) obtained by applying $1+\mathrm{B}+\mathrm{B}^2+\mathrm{B}^3$ and $1-\mathrm{B}$ successively to (Y_t). Show that (Z_t) is stationary and express its covariance function using the one of (X_t).
4. Characterize the spectral measure μ of (S_t).
5. Compute the spectral measure of $(1-\mathrm{B}^4)(Y_t)$ when (X_t) 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_t)?
2. What is the canonical representation of (X_t)?
3. What is the variance of the innovation process corresponding to (X_t)?
4. How is it possible to write X_t as a function of (ε_s) ?

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

$$Z_t = X_t + Y_t\,.$$

1. Show that (Z_t) 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_t).

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 $(\eta_t)_{t\in\mathbb{Z}}$ have variances σ_ε^2 and σ_η^2, respectively, and $0 < a, b < 1$. define

$$Z_t = X_t + Y_t\,.$$

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

Linear forecasting

In this chapter, we examine the problem of linear forecasting for time series. We first consider the case where the time series is a weakly stationary process. We then introduce a very general approach for modelling 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).

4.1 Linear forecasting for weakly stationary processes

4.1.1 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 2.4.2. 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

$$\Gamma_p^+ \phi_p^+ = \gamma_p^+ \ , \tag{4.1}$$

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

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

We are now interested in the effective computation of the prediction coefficients ϕ_p^+ (given γ) and of the prediction error defined by (2.18) and given by

$$\sigma_p^2 = \gamma(0) - (\phi_p^+)^H \gamma_p^+ , \tag{4.2}$$

see (2.19). The equations (4.1) and (4.2) 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 2.3.3 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 4.1.1. *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 (4.1) and (4.2). Then we have, for all $p = 0, 1, \dots,$*

$$\Gamma_{p+1}^+ = A_{p+1}^{-1} D_{p+1} (A_{p+1}^H)^{-1} , \tag{4.3}$$

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

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

By Definition 2.4.2, 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 $k \geq 1$, $\mathcal{H}_{k,k}^X = \operatorname{Span}(X_1, \dots, X_k)$ and thus increases with k. Using that $X_1 \in \mathcal{H}_{1,1}^X$ and for all $k = 2, \dots, p$, $\epsilon_{k,k-1}^+ \in \mathcal{H}_{k,k}^X$ and $\epsilon_{k+1,k}^+ \perp \mathcal{H}_{k,k}^X$, we get that $\begin{bmatrix} X_1 & \epsilon_{2,1}^+ & \dots & \epsilon_{p+1,p}^+ \end{bmatrix}^T$ have orthogonal components with variances $\sigma_0^2, \dots, \sigma_p^2$. Hence we obtain

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

from which we get (4.3). □

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 4.2).

Theorem 4.1.2. *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 (4.1) and (4.2). Then, if Γ_p^+ is invertible we have for all z in the closed unit disk $\{z \in \mathbb{C}, |z| \leq 1\}$,*

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

4.1.2 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 3: Levinson-Durbin algorithm.

Data: Covariance coefficients $\gamma(k)$, $k = 0, \ldots, K$
Result: Prediction coefficients $\{\phi_{m,p}^+\}_{1\leq m\leq p, 1\leq p\leq K}$, partial autocorrelation coefficients $\kappa(1), \ldots, \kappa(K)$
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 $p = 1, 2, \ldots, K-1$ **do**
Set

$$\kappa(p+1) = \sigma_p^{-2} \left(\gamma(p+1) - \sum_{k=1}^{p} \phi_{k,p}^+ \gamma(p+1-k) \right) \tag{4.4}$$

$$\sigma_{p+1}^2 = \sigma_p^2 (1 - |\kappa(p+1)|^2) \tag{4.5}$$

$$\phi_{p+1,p+1}^+ = \kappa(p+1) \tag{4.6}$$

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

$$\phi_{m,p+1}^+ = \phi_{m,p}^+ - \kappa(p+1)\, \overline{\phi_{p+1-m,p}^+} \, . \tag{4.7}$$

end
end

Observe that all the computations of Algorithm 3 can be done in $O(K^2)$ operations.

Theorem 4.1.3. *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 $p \geq 1$, (ϕ_p^+, σ_p^2) be any*

solution of the Yule-Walker equations (4.1) and (4.2). Then Algorithm 3 applies for any K such that Γ_K^+ is invertible, or, equivalently, $\sigma_{K-1}^2 > 0$.

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

Lemma 4.1.4. *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 $p \geq 1$, $\epsilon_{t,p}^+$ and $\kappa(p)$ are as in Definition 2.4.2 and Definition 3.5.1. Define moreover the* backward partial innovation process of order $p \geq 1$ *by*

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

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

$$\kappa(p+1) = \frac{\left\langle \epsilon_{t,p}^+, \epsilon_{t-p-1,p}^- \right\rangle}{\|\epsilon_{t,p}^+\| \, \|\epsilon_{t-p-1,p}^-\|} , \tag{4.8}$$

with the convention $0/0 = 0$.

Proof. Let us denote by c the right-hand side of (4.8) 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 $\Gamma_{p+1}^+ = \operatorname{Cov}\left(\begin{bmatrix} X_t & X_{t-1} & \dots & X_{t-p} \end{bmatrix}^T\right)$. 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 $\Gamma_{p+1}^- = \operatorname{Cov}\left(\begin{bmatrix} X_t & X_{t+1} & \dots & X_{t+p} \end{bmatrix}^T\right)$. 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 $\epsilon^-_{t-p-1,p} = X_{t-p-1} - \mathrm{proj}\left(X_{t-p-1} | \mathcal{H}^X_{t-1,p}\right)$. Hence,

$$c = \frac{\left\langle \epsilon^+_{t,p}, \epsilon^-_{t-p-1,p} \right\rangle}{\sigma_p^2} = \frac{\left\langle \epsilon^+_{t,p}, X_{t-p-1} \right\rangle}{\|\epsilon^+_{t-p-1,p}\|^2} = \frac{\left\langle X_t, \epsilon^-_{t-p-1,p} \right\rangle}{\|\epsilon^-_{t-p-1,p}\|^2} \ . \tag{4.9}$$

Moreover, we have

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

We thus get

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

and

$$\begin{aligned} &\|X_t - \mathrm{proj}\left(X_t | \mathcal{H}^X_{t-1,p+1}\right)\|^2 \\ &\quad = \|X_t - \mathrm{proj}\left(X_t | \mathcal{H}^X_{t-1,p}\right)\|^2 - \left\|\mathrm{proj}\left(X_t | \mathrm{Span}\left(\epsilon^-_{t-p-1,p}\right)\right)\right\|^2 \ . \end{aligned} \tag{4.10}$$

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

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

where we used (4.9). Moreover, by Theorem 4.1.1, $\sigma_p^2 \neq 0$ implies that Γ^+_p and Γ^+_{p+1} are invertible, so ϕ^+_p and ϕ^+_{p+1} are uniquely defined by (4.1) and the last two displays give

$$\sum_{k=1}^{p+1} \phi^+_{k,p+1} X_{t-k} = \sum_{k=1}^{p} \phi^+_{k,p} X_{t-k} + c \, \left(X_{t-p-1} - \sum_{k=1}^{p} \phi^-_{k,p} X_{t-p-1+k} \right) \ , \tag{4.11}$$

where $\phi^-_p = \left(\phi^-_{k,p}\right)_{k=1,\ldots,p}$ is uniquely defined by

$$\mathrm{proj}\left(X_{t-p-1} | \mathcal{H}^X_{t-1,p}\right) = \sum_{k=1}^{p} \phi^-_{k,p} X_{t-p-1+k} \ . \tag{4.12}$$

Since the prediction coefficients are uniquely defined in (4.11), we get by identifying those of the left-hand side with those of the right-hand side that

$$\phi^+_{k,p+1} = \phi^+_{k,p} - c\phi^-_{p+1-k,p} \qquad \text{for } k = 1, \ldots, p \tag{4.13}$$

$$\phi^+_{p+1,p+1} = c \tag{4.14}$$

Equation (4.14) gives (4.8), 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 4.1.1, 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 4.1.3 can now be completed.

Proof of Theorem 4.1.3. The initialization step is the usual projection formula in dimension 1.

We now prove the iteration formula, that is (4.4), (4.6), (4.5) and (4.7). Relation (4.6) is proved in Lemma 4.1.4. Under the assumptions of Theorem 4.1.3, we can use the facts shown in the proof of Lemma 4.1.4 in the case where $\sigma_p^2 \neq 0$. Relation (4.9) 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 (4.4).

Relation (4.10) implies that

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

that is, by definition of c, we get (4.5).

To prove (4.7) with (4.13), we need to relate ϕ_p^- (uniquely defined by (4.12) with ϕ_p^+, solution of (4.1). Similarly, 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

$$\Gamma_p^- = \mathrm{Cov}\left([X_{t-p} \ \dots \ X_{t-1}]^T \right)^T \tag{4.15}$$

$$= \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & & \gamma(p-1) \\ \gamma(-1) & \gamma(0) & \gamma(1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(1) \\ \gamma(1-p) & \gamma(2-p) & \cdots & \gamma(-1) & \gamma(0) \end{bmatrix} . \tag{4.16}$$

Hence $\gamma_p^- = \overline{\gamma_p^+}$ and $\Gamma_p^- = \overline{\Gamma_p^+}$ and so $\phi_p^- = \overline{\phi_p^+}$. This, with (4.13), yields (4.7), which concludes the proof. □

4.1.3 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 4.1.1. 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 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

$$\gamma(j,k) = \mathrm{Cov}\,(X_j, X_k)\,, \quad j,k \geq 1\,. \tag{4.17}$$

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

$$\epsilon_1 = X_1 \quad \text{and} \quad \epsilon_t = X_t - \mathrm{proj}\left(X_t|\,\mathcal{H}_{t-1}^X\right), \quad t = 2,3,\ldots \tag{4.18}$$

As in the usual the Gram-Schmidt procedure, we immediately obtain that $(\epsilon_t)_{n\in\mathbb{N}}\, t \geq 1$ is an orthogonal sequence, moreover we have, for all $p \geq 1$,

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

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 4: 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 4 applies also in the case where X is weakly stationary. Observe that all the computations of Section 4.1.3 can be done in $O(K^3)$ operations. Hence in the weakly stationary case, one should prefer Algorithm 3 to Algorithm 4. On the other hand, there is one case where Algorithm 4 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 4.1.1 (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 4*

boils down to

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

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

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

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

4.2 Exercises

Exercise 4.1 (Linear prediction of an AR(1) observed with additive noise). Consider an AR(1) process Z_t satisfying the following canonical equation :

$$Z_{t+1} = \phi Z_t + \eta_t \quad \text{for } t \in \mathbb{Z} \tag{4.19}$$

where $(\eta_t)_{t\geq 0}$ is a centered white noise with known variance σ^2 and ϕ is a known constant. The process $(Z_t)_{t\geq 0}$ is not directly observed. Instead, for all $t \geq 1$, one gets the following sequence of observations :

$$Y_t = Z_t + \varepsilon_t \tag{4.20}$$

where $(\varepsilon_t)_{t\geq 1}$ is a centered white noise with known variance ρ^2, that is uncorrelated with (η_t) and Z_0. We wish to solve the filtering problem, that is, to compute the orthogonal projection of Z_t on the space $H_t = \text{span}\{Y_1, \cdots, Y_t\}$, *iteratively in* t.

We denote $\hat{Z}_{t|t} = \text{proj}\,(Z_t | H_t)$ this projection and $P_{t|t} = \mathbb{E}\left[(Z_t - \hat{Z}_{t|t})^2\right]$ the corresponding projection error variance. Similarly, let $\hat{Z}_{t+1|t} = \text{proj}\,(Z_{t+1} | H_t)$ be the best linear predictor and $P_{t+1|t} = \mathbb{E}(Z_{t+1} - \hat{Z}_{t+1|t})^2$ the linear prediction error variance.

1. Show that Z_0 is a centered random variable and computes its variance σ_0^2 and that Z_0 and $(\eta_t)_{t\geq 0}$ are uncorrelated.
2. Using the evolution equation (4.19), show that
$$\hat{Z}_{t+1|t} = \phi \hat{Z}_{t|t} \quad \text{et} \quad P_{t+1|t} = \phi^2 P_{t|t} + \sigma^2$$
3. Let us define the innovation by $I_{t+1} = Y_{t+1} - \text{proj}\,(Y_{t+1} | H_t)$. Using the observation equation (4.20), show that $I_{t+1} = Y_{t+1} - \hat{Z}_{t+1|t}$.
4. Prove that $\mathbb{E}[I_{t+1}^2] = P_{t+1|t} + \rho^2$.
5. Give the argument that shows that
$$\hat{Z}_{t+1|t+1} = \hat{Z}_{t+1|t} + k_{t+1} I_{t+1}$$
where $k_{t+1} = \mathbb{E}[Z_{t+1} I_{t+1}]/\mathbb{E}[I_{t+1}^2]$.
6. Using the above expression of I_{t+1}, show that $\mathbb{E}[Z_{t+1} I_{t+1}] = P_{t+1|t}$.
7. Why is the following equation correct ?
$$P_{t+1|t+1} = P_{t+1|t} - \mathbb{E}\left[(k_{t+1} I_{t+1})^2\right]$$
Deduce that $P_{t+1|t+1} = (1 - k_{t+1}) P_{t+1|t}$.

8. Provide the complete set of equations for computing $\hat{Z}_{t|t}$ and $P_{t|t}$ iteratively for all $t \geq 1$. (Include the initial conditions.)

9. Study the asymptotic behavior of $P_{t|t}$ as $t \to \infty$.

Exercise 4.2. Let $(X_t)_{t\in\mathbb{Z}}$ and, (ϕ_p^+, σ_p^2), $p \geq 1$ be as in Theorem 4.1.2.

1. Compute (ϕ_1^+, σ_1^2) 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 4.1.2.

Exercise 4.3. Show that the process $(Y_t)_{t\in\mathbb{Z}}$ of Exercise 4.1 is an ARMA(1,1) process if $|\phi| < 1$ and σ_0^2 is set to a well chosen value.

Chapter 5

Kalman filter

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.

5.1 Conditional mean for Gaussian vectors

Let $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P})$, which is an Hilbert space. Let $\mathcal{G}$ be σ-field included in $\mathcal{F}$ and $\mathcal{E} = L^2(\Omega, \mathcal{G}, \mathbb{P})$. For any $X \in \mathcal{H}$, one can define the *conditional expectation* of X given $\mathcal{E}$ by:

$$\mathbb{E}\left[X \middle| \mathcal{G}\right] = \mathrm{proj}\left(X \middle| \mathcal{E}\right) \ .$$

If $\mathcal{G}$ is generated by a collection of random variables, say $\mathcal{G} = \sigma(\mathbf{Y})$, we denote

$$\mathbb{E}\left[X \middle| \mathbf{Y}\right] = \mathbb{E}\left[X \middle| \mathcal{G}\right] \ .$$

And, as for standard expectation, if $\mathbf{X}$ is a random vector, $\mathbb{E}\left[\mathbf{X} \middle| \mathcal{G}\right]$ is the vector made of the conditional expectations of its entries.

In the Gaussian context, the following result moreover holds, whose proof is left as an exercise (see Exercise 5.1).

Proposition 5.1.1. *Let $p, q \geq 1$. Let $\mathbf{X}$ and $\mathbf{Y}$ be two jointly Gaussian vectors, respectively valued in $\mathbb{R}^p$ and $\mathbb{R}^q$. Let*

$$\widehat{\mathbf{X}} = \mathrm{proj}\left(\mathbf{X} \middle| \mathrm{Span}\left(1, \mathbf{Y}\right)\right) \ ,$$

where here $\mathrm{Span}\left(\dots\right)$ is understood as the space of $\mathbb{R}^p$-valued L^2 random variables obtained by linear transformations of $\dots$ and $\mathrm{proj}\left(\cdot \middle| \dots\right)$ is understood as the projection onto this space seen as a (closed) subspace of the Hilbert space of all $\mathbb{R}^p$-valued L^2 random variables. Then the following assertions hold.

(i) We have

$$\mathrm{Cov}(\mathbf{X}-\widehat{\mathbf{X}}) = \mathbb{E}\left[\mathbf{X}(\mathbf{X}-\widehat{\mathbf{X}})^T\right] = \mathbb{E}\left[(\mathbf{X}-\widehat{\mathbf{X}})\mathbf{X}^T\right] .$$

(ii) We have

$$\mathbb{E}\left[\mathbf{X}|\,\mathbf{Y}\right] = \mathrm{proj}\left(\mathbf{X}|\,\mathrm{Span}\left(1,\mathbf{Y}\right)\right) ,$$

(iii) If moreover $\mathrm{Cov}(\mathbf{Y})$ *is invertible, then*

$$\widehat{\mathbf{X}} = \mathbb{E}\left[\mathbf{X}\right] + \mathrm{Cov}\left(\mathbf{X},\mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\left(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}\right]\right) ,$$

and

$$\mathrm{Cov}(\mathbf{X}-\widehat{\mathbf{X}}) = \mathrm{Cov}(\mathbf{X}) - \mathrm{Cov}\left(\mathbf{X},\mathbf{Y}\right)\mathrm{Cov}(\mathbf{Y})^{-1}\mathrm{Cov}\left(\mathbf{Y},\mathbf{X}\right) .$$

5.2 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 [7, 8] 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 5.2.1 (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 5.2.1 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 5.2.1. $(\mathbf{X}_t)_{t\geq 1}$ *and* $(\mathbf{Y}_t)_{t\geq 1}$ *are* p*-dimensional and* q*-dimensional time series satisfying the following equations for all* $t \geq 1$*,*

$$\mathbf{X}_t = \Phi_t\mathbf{X}_{t-1} + \mathrm{A}_t\mathbf{u}_t + \mathbf{W}_t , \tag{5.1}$$
$$\mathbf{Y}_t = \Psi_t\mathbf{X}_t + \mathrm{B}_t\mathbf{u}_t + \mathbf{V}_t , \tag{5.2}$$

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 5.1.1, 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 of $\mathbf{X}$ on the space of linear combinations of $\mathbf{Y}$ 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 $\mathbb{E}[\mathbf{X}|\,\mathbf{Y}]$ are replaced by $\mathrm{proj}(\mathbf{X}|\,\mathrm{Span}(1, \mathbf{Y}))$, see Corollary 5.3.3.

Remark 5.2.1. *A slight extension of this model is to let the covariance matrices* R *and* Q *depend on* t*. All the results of Section Section 5.3 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* (5.1) 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* (5.2) 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 5.2.1 (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 $Y_t = [1 \quad 0]X_t + V_t$ *The process* (Z_t) *is obtained from* (X_t) *by* $Z_t = [1 \quad 0]X_t$. *We display a simulated* (Z_t) *and* (Y_t) *in Figure Figure 5.1.*

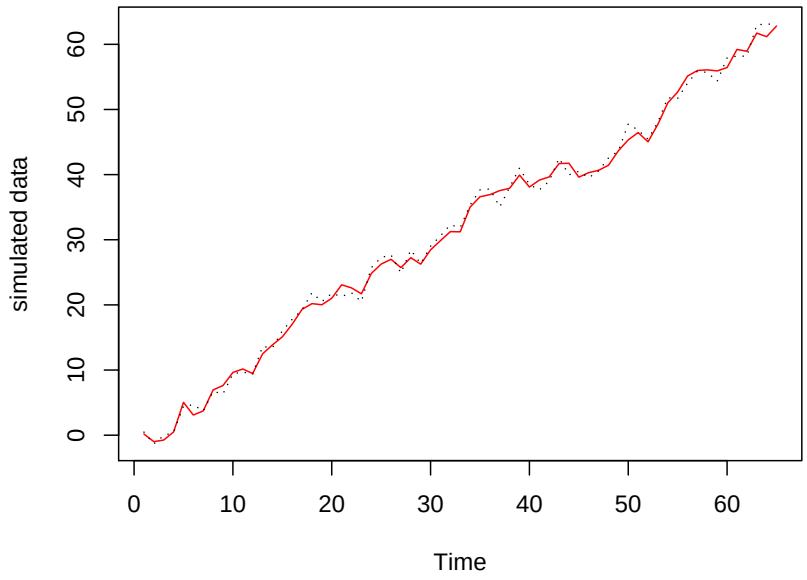

Figure 5.1: Simulated random trend (plain red line) and its observation with additive noise (dotted black line).

Example 5.2.2 (Climatology data)**.** *Figure 5.2 shows two different estimates of the global temperature deviations from 1880 to 2009. They can be found on the site*

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

$$R = \mathrm{Cov} \begin{bmatrix} V_{1,t} \\ V_{2,t} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}.$$

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 $\mathrm{B}_t = 0$.

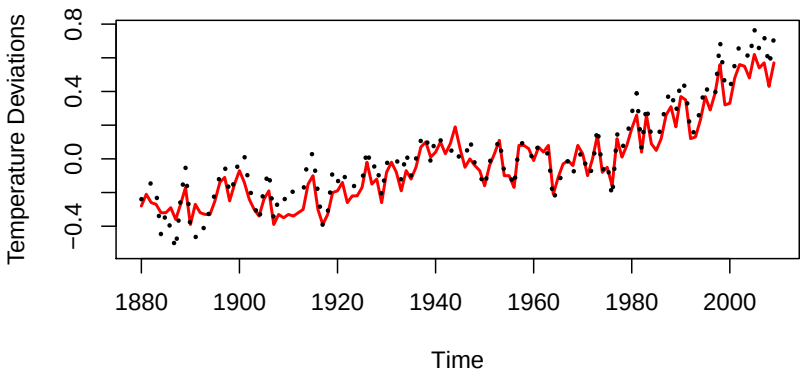

Figure 5.2: Annual global temperature deviation series, measured in degrees centigrade, 1880–2009.

Dynamic linear models allow us to provide a quite general framework for denoising and forecasting a Gaussian process, or/and estimating its parameters. In (5.1) and (5.2), 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 5.2.3 (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 5.3. We easily compute*

$$\gamma_y(0) = \text{Var}\,(Y_t) = \text{Var}\,(X_t + V_t) = \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2, \tag{5.3}$$

and, when $h \neq 0$,

$$\gamma_y(h) = \text{Cov}\,(Y_t, Y_{t-h}) = \text{Cov}\,(X_t + V_t, X_{t-h} + V_{t-h}) = \gamma_x(h).$$

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 5.3). We will provide a general view on the relationships between DLMs and stationary ARMA processes in Section 5.6.*

5.3 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 $\mathbf{Y}_{1:s} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_s\}$. More precisely, it consists in computing the conditional mean

$$\mathbf{X}_{t|s} \overset{\text{def}}{=} \mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{Y}_{1:s}\right] \tag{5.4}$$

and to measure the L^2 norm of the error $\mathbf{X}_t - \mathbf{X}_{t|s}$,

$$\mathbf{\Sigma}_{t|s} \overset{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_t - \mathbf{X}_{t|s})(\mathbf{X}_t - \mathbf{X}_{t|s})^T\right] = \text{Cov}(\mathbf{X}_t - \mathbf{X}_{t|s})\ , \tag{5.5}$$

since $\mathbf{X}_t - \mathbf{X}_{t|s}$ is centered.

Three different situations are generally distinguished.

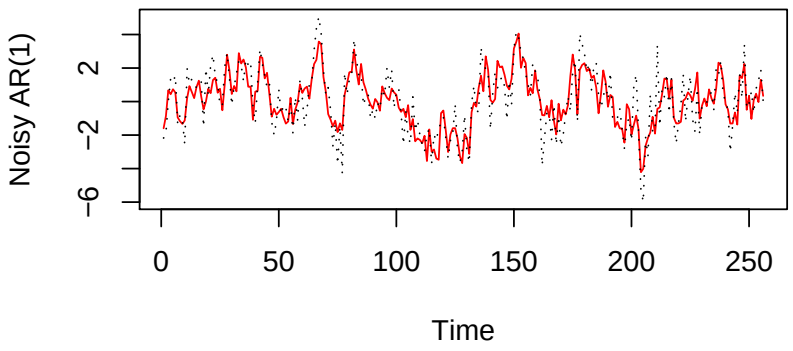

Figure 5.3: Simulated AR(1) process (solid red) and a noisy observation of it (dotted black).

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 5.7.

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 5: 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 $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$ for $t = 1, \ldots, n$.

Initialization: set $\mathbf{X}_{0|0} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{0|0} = \Sigma_0$.

for $t = 1, 2, \ldots, n$ **do**

Compute in this order

$$\mathbf{X}_{t|t-1} = \Phi_t \mathbf{X}_{t-1|t-1} + \mathrm{A}_t \mathbf{u}_t, \tag{5.6}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \Phi_t \boldsymbol{\Sigma}_{t-1|t-1} \Phi_t^T + Q, \tag{5.7}$$

$$K_t = \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T [\Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R]^{-1}. \tag{5.8}$$

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + K_t(\mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t), \tag{5.9}$$

$$\boldsymbol{\Sigma}_{t|t} = [I - K_t \Psi_t] \boldsymbol{\Sigma}_{t|t-1}. \tag{5.10}$$

end

Proposition 5.3.1 (Kalman Filter)**.** *Algorithm 5 holds for the state-space model satisfying Assumption 5.2.1, provided that $\Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R$ are invertible matrices for $t = 1, \ldots, n$.*

The matrix K_t defined in (5.8) is called the *Kalman gain matrix*.

For proving Proposition 5.3.1, we will introduce the following definition

$$\begin{aligned} \mathbf{Y}_{t|s} &\stackrel{\text{def}}{=} \mathbb{E}\left[\mathbf{Y}_t \middle| \mathbf{Y}_{1:s}\right] \\ \boldsymbol{\epsilon}_t &\stackrel{\text{def}}{=} \mathbf{Y}_t - \mathbf{Y}_{t|t-1} \\ \Gamma_t &\stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T\right] = \mathrm{Cov}(\boldsymbol{\epsilon}_t)\ , \end{aligned}$$

and show the folowing useful formula

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t, \tag{5.11}$$

$$\Gamma_t = \mathrm{Cov}(\Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t) = \Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R \tag{5.12}$$

for $t = 1, \ldots, n$. The process $(\boldsymbol{\epsilon}_t)$ is called the *innovation process* of $(\mathbf{Y}_t)$.

Proof of Proposition 5.3.1. By Assumption 5.2.1, 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 (5.1), this implies that, for all $t > s$

$$\begin{aligned} \mathbf{X}_{t|s} &= \mathbb{E}\left[\mathbf{X}_t \middle| \mathbf{Y}_{1:s}\right] \\ &= \mathbb{E}\left[\Phi_t \mathbf{X}_{t-1} + \mathrm{A}_t \mathbf{u}_t + \mathbf{W}_t \middle| \mathbf{Y}_{1:s}\right] \\ &= \Phi_t \mathbf{X}_{t-1|s} + \mathrm{A}_t \mathbf{u}_t\ , \end{aligned} \tag{5.13}$$

and, moreover,

$$\begin{aligned}\boldsymbol{\Sigma}_{t|s} &= \mathrm{Cov}\left(\mathbf{X}_t - \mathbf{X}_{t|s}\right) \\ &= \mathrm{Cov}\left(\Phi_t(\mathbf{X}_{t-1} - \mathbf{X}_{t-1|s}) + \mathbf{W}_t\right) \\ &= \Phi_t \boldsymbol{\Sigma}_{t-1|s} \Phi_t^T + Q \ . \end{aligned} \tag{5.14}$$

which gives (5.6) and (5.7).

Next, we show (5.9). 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 5.2.1, the variables $\mathbf{X}_t$, $\mathbf{Y}_{1:t-1}$ and $\boldsymbol{\epsilon}_t$ are jointly Gaussian. It then follows from Proposition Proposition 5.1.1 that

$$\mathbf{X}_{t|t} = \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(1, \mathbf{Y}_{1:t-1}, \boldsymbol{\epsilon}_t\right)\right) \ ,$$

where here $\mathrm{Span}\left(\dots\right)$ is understood as the space of $\mathbb{R}^p$-valued L^2 random variables obtained by linear transformations of $\dots$ and $\mathrm{proj}\left(\cdot|\dots\right)$ is understood as the projection onto this space seen as a (closed) subspace of the Hilbert space of all $\mathbb{R}^p$-valued L^2 random variables. Observing that $\boldsymbol{\epsilon}_t$ is centered and uncorrelated with $\mathbf{Y}_{1:t-1}$, we further have

$$\mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(1, \mathbf{Y}_{1:t-1}, \boldsymbol{\epsilon}_t\right)\right) = \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(1, \mathbf{Y}_{1:t-1}\right)\right) + \mathrm{proj}\left(\mathbf{X}_t \middle| \mathrm{Span}\left(\boldsymbol{\epsilon}_t\right)\right)$$

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

$$\begin{aligned}\boldsymbol{\Sigma}_{t|t} &= \boldsymbol{\Sigma}_{t|t-1} - \mathrm{Cov}(K_t \boldsymbol{\epsilon}_t) \\ &= \boldsymbol{\Sigma}_{t|t-1} - K_t \Gamma_t K_t^T \\ &= \boldsymbol{\Sigma}_{t|t-1} - K_t \mathrm{Cov}\left(\mathbf{X}_t, \mathbf{Y}_t - \mathbf{Y}_{t|t-1}\right)^T \ . \end{aligned}$$

Now, by (5.2), 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

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{X}_t, \mathbf{Y}_t - \mathbf{Y}_{t|t-1}\right) &= \mathrm{Cov}\left(\mathbf{X}_t, \Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t\right) \\ &= \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T \ , \end{aligned}$$

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 (5.8), (5.9) and (5.10) 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 5 in the sense that in practice one can use

them after having first applied Algorithm 5.

Algorithm 6: Kalman forecasting algorithm.

Data: A forecasting lag h, parameters Q and A_t for $t = n+1, \dots, n+h$, and exogenous input series $\mathbf{u}_t$, for $t = n+1, \dots, n+h$, 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 $t = n+1, \dots, n+h$

Initialization: set $k = 1$.

for $k = 1, 2, \dots, h$ **do**

Compute in this order

$$\mathbf{X}_{n+k|n} = \Phi_{n+k}\mathbf{X}_{n+k-1|n} + \mathrm{A}_{k+n}\mathbf{u}_{n+k} \,,$$
$$\mathbf{\Sigma}_{t|s} = \Phi_{n+k}\mathbf{\Sigma}_{t-1|s}\Phi_{n+k}^T + Q \,.$$

end

Algorithm 7: Rauch-Tung-Striebel smoother algorithm.

Data: Parameters Φ_t for $t = 1, \dots, n$, and exogenous input series $\mathbf{u}_t$, for $t = n+1, \dots, n+h$, 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 $t = 1, \dots, n$.

Result: Smoothing outputs $\mathbf{X}_{t|n}$, and their autocovariance matrices $\mathbf{\Sigma}_{t|n}$ for $t = n-1, n-2 \dots, 1$.

for $t = n, n-1, \dots, 2$ **do**

Compute in this order

$$J_{t-1} = \mathbf{\Sigma}_{t-1|t-1}\Phi_t^T\mathbf{\Sigma}_{t|t-1}^{-1} \,, \tag{5.15}$$
$$\mathbf{X}_{t-1|n} = \mathbf{X}_{t-1|t-1} + J_{t-1}\left(\mathbf{X}_{t|n} - \mathbf{X}_{t|t-1}\right), \tag{5.16}$$
$$\mathbf{\Sigma}_{t-1|n} = \mathbf{\Sigma}_{t-1|t-1} + J_{t-1}\left(\mathbf{\Sigma}_{t|n} - \mathbf{\Sigma}_{t|t-1}\right)J_{t-1}^T \,. \tag{5.17}$$

end

Proposition 5.3.2. *Algorithm 6 and Algorithm 7 hold for the state-space model satisfying Assumption 5.2.1, provided that (only for Algorithm 7)* $\mathbf{\Sigma}_{t|t-1}$ *is an invertible matrix for* $t = 2, \dots, n$.

Proof. Algorithm 6 directly follows from (5.13) and (5.14).

We now show that Algorithm 7 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

$$\mathbb{E}\left[\mathbf{X}_{t-1} \middle| \mathbf{Y}_{1:n}\right] = \mathbb{E}\left[\widetilde{\mathbf{X}}_{t-1} \middle| \mathbf{Y}_{1:n}\right] \,, \tag{5.18}$$

where

$$\begin{aligned}\widetilde{\mathbf{X}}_{t-1} &= \mathbb{E}\left[\mathbf{X}_{t-1} | \, \mathbf{Y}_{1:t-1}, \, \mathbf{X}_t - \mathbf{X}_{t|t-1}, \, \mathbf{V}_{t:n}, \mathbf{W}_{t+1:n}\right] \\ &= \mathbb{E}\left[\mathbf{X}_{t-1} | \, \mathbf{Y}_{1:t-1}, \, \mathbf{X}_t - \mathbf{X}_{t|t-1}\right] \, ,\end{aligned}$$

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

$$\widetilde{\mathbf{X}}_{t-1} = \mathbf{X}_{t-1|t-1} + J_{t-1}(\mathbf{X}_t - \mathbf{X}_{t|t-1}), \tag{5.19}$$

and

$$\mathrm{Cov}\left(\mathbf{X}_{t-1} - \widetilde{\mathbf{X}}_{t-1}\right) = \mathbf{\Sigma}_{t-1|t-1} - J_{t-1}\mathbf{\Sigma}_{t|t-1}J_{t-1}^T, \tag{5.20}$$

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 (5.15). By (5.18) and (5.19), we obtain, by projecting $\widetilde{\mathbf{X}}_{t-1}$ on $\mathrm{Span}\,(1, \mathbf{Y}_{1:n})$,

$$\mathbf{X}_{t-1|n} = \mathbf{X}_{t-1|t-1} + J_{t-1}(\mathbf{X}_{t|n} - \mathbf{X}_{t|t-1}) \, ,$$

that is (5.16) 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 (5.20), 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 (5.17). □

Inspecting the proofs of Proposition 5.3.1 and Proposition 5.3.2, 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 5.3.3. *Suppose that Assumption 5.2.1 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 5, Algorithm 6 and Algorithm 7 continue to hold if the definitions of* $\mathbf{X}_{s|t}$ *in (5.6) is replaced by*

$$\mathbf{X}_{s|t} \overset{\text{def}}{=} \mathrm{proj}\left(\mathbf{X}_t | \, \mathrm{Span}\,(1, \mathbf{Y}_{1:t-1})\right) \, ,$$

where here $\mathrm{Span}\,(\dots)$ *is understood as the space of* $\mathbb{R}^p$*-valued* L^2 *random variables obtained by linear transformations of* $\dots$ *and* $\mathrm{proj}\,(\cdot|\dots)$ *is understood as the projection onto this space seen as a (closed) subspace of the Hilbert space of all* $\mathbb{R}^p$*-valued* L^2 *random variables.*

For estimation purposes, we will need to compute the one-lag covariance matrix of the smoother outputs that is

$$\mathbf{\Sigma}_{t_1,t_2|s} \stackrel{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_{t_1} - \mathbf{X}_{t_1|s})(\mathbf{X}_{t_2} - \mathbf{X}_{t_2|s})^T\right] \tag{5.21}$$

with $t_1 = t, t_2 = t-1$ and $s = n$. Note that this notation extends the previous one in the sense that $\mathbf{\Sigma}_{t|s} = \mathbf{\Sigma}_{t,t|s}$.

One simple way to compute $\mathbf{\Sigma}_{t-1,t-2|n}$ is to define new state and observation variables by stacking two consecutive times together, namely

$$\mathbf{X}_{(t)} \stackrel{\text{def}}{=} [\mathbf{X}_t^T \ \mathbf{X}_{t-1}^T]^T,$$
$$\mathbf{Y}_{(t)} \stackrel{\text{def}}{=} [\mathbf{Y}_t^T \ \mathbf{Y}_{t-1}^T]^T.$$

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

$$\mathbf{\Sigma}_{(t)|(n)} = \begin{bmatrix} \mathbf{\Sigma}_{t|n} & \mathbf{\Sigma}_{t|t-1}n \\ \mathbf{\Sigma}_{t,t-1|n}^T & \mathbf{\Sigma}_{t-1|n} \end{bmatrix},$$

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 5.2).

Algorithm 8: One-lag covariance algorithm.

Data: Parameters Ψ_n and Φ_t for $t = 1, \dots, n$, Gain matrix K_n Kalman filter covariance matrices $\mathbf{\Sigma}_{t|t}$ and $\mathbf{\Sigma}_{t|t-1}$ for $t = 1, \dots, n$, matrices J_t for $t = 1, \dots, n-1$.

Result: One-lag covariance matrices for smoother outputs $\mathbf{\Sigma}_{t,t-1|n}$ for $t = 1, \dots, n$.

Initialization: Set

$$\mathbf{\Sigma}_{n,n-1|n} = (I - K_n\Psi_n)\Phi_n\mathbf{\Sigma}_{n-1|n-1}, \tag{5.22}$$

for $t = n, n-1, \dots, 2$ **do**

$$\mathbf{\Sigma}_{t-1,t-2|n} = \mathbf{\Sigma}_{t-1|t-1}J_{t-2}^T + J_{t-1}\left(\mathbf{\Sigma}_{t,t-1|n} - \Phi_t\mathbf{\Sigma}_{t-1|t-1}\right)J_{t-2}^T. \tag{5.23}$$

end

Remark 5.3.1. *All the above algorithms (Algorithms 5, 6, 7 and 8) are* recursive *in the sense that their outputs are computed using a simple recursive set of equations. Algorithm 5 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 5.3.2. *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 5.3.1 (Noisy AR(1) (continued from Example 5.2.3))**.** *Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ be as in Example 5.2.3. 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 (5.3)). To produce Figure 5.4, 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 $\mathcal{N}(\mathbf{X}_{t|n}, \mathbf{\Sigma}_{t|n})$.*

5.4 Steady State approximations

Let us consider Assumption 5.2.1 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 5, 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

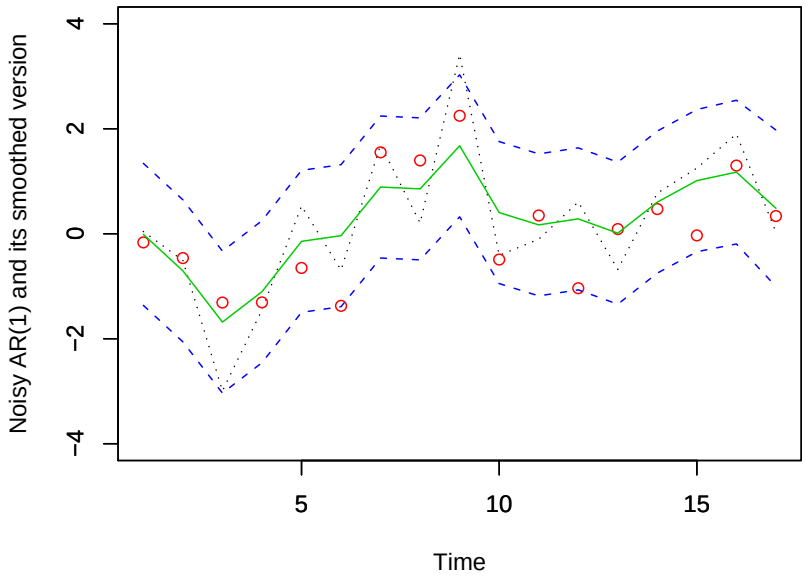

Figure 5.4: Simulated AR(1) process (red circles), a noisy observation of it (dotted black line), the smoother outputs (solid green line) and the 95% confidence intervals (between blued dashed lines).

$\mathbf{\Sigma}_{t|t-1}$ and $\mathbf{\Sigma}_{t-1|t-2}$ (following (5.10) and (5.10)), we obtain that Σ is a necessary solution to the *Ricatti equation*

$$\Sigma = \Phi[\Sigma - \Sigma\Psi^T(\Psi\Sigma\Psi^T + R)^{-1}\Psi\Sigma]\Phi^T + Q \,, \tag{5.24}$$

and following (5.8), 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 (6.37), can be established when Φ has eigenvalues within the open unit disk $\{z \in \mathbb{C}\,, |z| < 1\}$. We just refer to [4, 5] for details. Let us just briefly give a hint of why this assumption is meaningful. Iterating the state equation (5.1) 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 6.2, one can derive the asymptotic behavior of the MLE, under appropriate assumptions of the parameterization.

5.5 Correlated Errors

Sometimes it is advantageous to use assumptions for the linear state-space model which are slightly different from Assumption 5.2.1. 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 5.5.1. *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 $t \geq 1$,*

$$\mathbf{X}_{t+1} = \Phi_t \mathbf{X}_t + \mathrm{A}_{t+1}\mathbf{u}_{t+1} + \Theta_t \mathbf{W}_t \ , \tag{5.25}$$

$$\mathbf{Y}_t = \Psi_t \mathbf{X}_t + \mathrm{B}_t \mathbf{u}_t + \mathbf{V}_t \ , \tag{5.26}$$

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 5 has to be adapted as follows.

Algorithm 9: 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

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t \tag{5.27}$$

$$\Gamma_t = \Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R, \tag{5.28}$$

$$K_t = [\Phi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + \Theta_t S] \Gamma_t^{-1}, \tag{5.29}$$

$$\mathbf{X}_{t+1|t} = \Phi_t \mathbf{X}_{t|t-1} + \mathrm{A}_{t+1} \mathbf{u}_{t+1} + K_t \boldsymbol{\epsilon}_t, \tag{5.30}$$

$$\boldsymbol{\Sigma}_{t+1|t} = \Phi_t \boldsymbol{\Sigma}_{t|t-1} \Phi_t^T + \Theta_t Q \Theta_t^T - K_t \Gamma_t^{-1} K_t^T, \tag{5.31}$$

$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T \Gamma_t^{-1} \boldsymbol{\epsilon}_t, \tag{5.32}$$

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1} \Psi_{t+1}^T \Gamma_t^{-1} \Psi_t \boldsymbol{\Sigma}_{t|t-1}. \tag{5.33}$$

end

In this algorithm, $\boldsymbol{\epsilon}_t$ and Γ_t still correspond to the innovation process and its covariance matrix,

$$\begin{aligned} \boldsymbol{\epsilon}_t &\stackrel{\text{def}}{=} \mathbf{Y}_t - \mathbf{Y}_{t|t-1} \\ \Gamma_t &\stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T\right] = \mathrm{Cov}(\boldsymbol{\epsilon}_t) \ . \end{aligned}$$

The adaptation of the proof of Proposition 5.3.1 to the correlated errors case is left to the reader (Exercise 5.4). The following result follows.

Proposition 5.5.1 (Kalman Filter for correlated errors)**.** *Algorithm 9 applies for the state-space model satisfying Assumption 5.5.1, provided that* $\Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R$ *are invertible matrices for* $t = 1, \ldots, n$.

5.6 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

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \sum_{j=1}^{p} \Phi_j \mathbf{Y}_{t-j} + \sum_{k=1}^{q} \Theta_k \mathbf{V}_{t-k} + \mathbf{V}_t \ . \tag{5.34}$$

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 5.5.1, under the additional Gaussian assumption. The proof is left to the reader (Exercise 5.5).

Proposition 5.6.1 (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*

$$\mathbf{X}_{t+1} = \Phi \mathbf{X}_t + \mathrm{A}\mathbf{u}_{t+1} + \Theta \mathbf{V}_t, \tag{5.35}$$

$$\mathbf{Y}_t = \Psi \mathbf{X}_t + \mathbf{V}_t, \tag{5.36}$$

where $\Psi = \begin{bmatrix} \mathbf{1}, 0, \cdots, 0 \end{bmatrix}$ is $k \times pk$, implies the ARMAX model (5.34). If $p < q$, set $\Phi_{p+1} = \cdots = \Phi_q = 0$, and replace the value of p by that of q and (5.35)–(5.36) still apply. Note that the state process is kp-dimensional, whereas the observations are k-dimensional.

Example 5.6.1 (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 5.6.1, we can write the model as

$$X_{t+1} = \phi X_t + \beta_0 + \beta_1 t + (\theta + \phi) V_t, \tag{5.37}$$

and

$$Y_t = X_t + V_t. \tag{5.38}$$

Remark 5.6.1. *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 5.6.2 (Regression with autocorrelated errors)**.** *The (multivariate) regression with autocorrelated errors, is the regression model*

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t \,, \tag{5.39}$$

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*

$$\mathbf{X}_{t+1} = \Phi\mathbf{X}_t + \Theta\mathbf{V}_t, \tag{5.40}$$

$$\mathbf{Y}_t = \Psi\mathbf{X}_t + \mathrm{B}\mathbf{u}_t + \mathbf{V}_t, \tag{5.41}$$

where the model matrices Φ*,* Θ*, and* Ψ *are defined in Proposition 5.6.1.*

5.7 Likelihood of dynamic linear models

The dynamic linear model of Assumption 5.2.1 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 5.2.1 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 5.3 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 (5.11) and (5.12) become

$$\boldsymbol{\epsilon}_t(\theta) = \mathbf{Y}_t - \Psi_t(\theta)\mathbf{X}_{t|t-1}(\theta) - \mathrm{B}_t(\theta)\mathbf{u}_t, \tag{5.42}$$

$$\Gamma_t(\theta) = \Psi(\theta)\boldsymbol{\Sigma}_{t|t-1}(\theta)\Psi_t(\theta)^T + R(\theta) \tag{5.43}$$

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 (6.35), (6.38) and (6.39), in the framework of a parameterized dynamic linear model, (6.40) thus gives that

$$-2\log L_n(\theta) = n\log(2\pi) + \sum_{t=1}^{n} \log\det\Gamma_t(\theta) + \sum_{t=1}^{n} \boldsymbol{\epsilon}_t(\theta)^T \Gamma_t(\theta)^{-1} \boldsymbol{\epsilon}_t(\theta)\ , \tag{5.44}$$

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 5) and then applying (5.42), (5.43) and (5.44).

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 (6.41) and (6.42) can directly be applied replacing $\boldsymbol{\eta}$ by $\boldsymbol{\epsilon}$ and $\widetilde{\Sigma}$ by Γ.

However one needs to adapt Algorithm 5 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 5 provides the following algorithm.

Algorithm 10: 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 5 and Algorithm 10 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 5.3.1, 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 5.7.1 (Noisy AR(1) (continued from Example 5.2.3 and Example 5.3.1)**.** *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 5.2.3 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 5.5.*

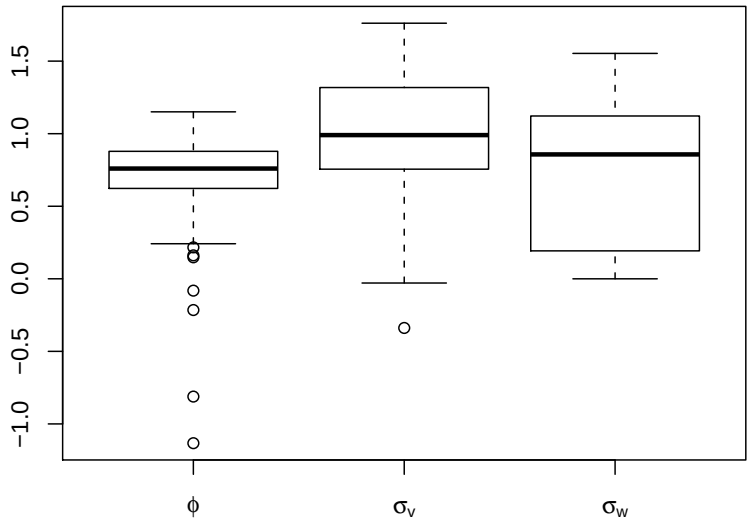

Figure 5.5: Estimation of the parameters of the noisy AR(1) model: boxplots of the estimates of ϕ, σ_v and σ_w obtained from 100 Monte Carlo replications of time series of length 128. The true values are $\phi = 0.8$ and $\sigma_v = 1.0$ and $\sigma_w = 1.0$.

[1] the quasi Newton procedure implemented in the `optim()` function of the R software, [10]

5.8 Exercises

Exercise 5.1. Let $\mathbf{X}$, $\mathbf{Y}$ and $\widehat{\mathbf{X}}$ be as in Proposition 5.1.1. Let $\boldsymbol{\epsilon} = \mathbf{X} - \widehat{\mathbf{X}}$.

1. Explain why there exists $\mathbf{a} \in \mathbb{R}^q$ and $A \in \mathbb{R}^{p\times q}$ such that $\widehat{\mathbf{X}} = \mathbf{a} + A\mathbf{Y}$ and for all $\mathbf{a}' \in \mathbb{R}^q$ and $A' \in \mathbb{R}^{p\times q}$ we have
$$\mathbb{E}\left[\boldsymbol{\epsilon}^T\left(\mathbf{a}' + A'\mathbf{Y}\right)\right] = 0\,.$$
2. Show that $\mathbb{E}[\boldsymbol{\epsilon}] = 0$.
3. Show that Proposition 5.1.1(i) holds.
4. Show that $\boldsymbol{\epsilon}$ and $\mathbf{Y}$ are jointly Gaussian and independent. Deduce that Proposition 5.1.1(ii) holds.
5. Show that Proposition 5.1.1(iii) holds.

Exercise 5.2. Show that Algorithm 8 applies under the assumptions of Proposition 5.3.1.

Exercise 5.3. Show that the process $(Y_t)_{t\in\mathbb{Z}}$ of Example 5.2.3 is an ARMA(1,1) process.

Exercise 5.4. Show that Algorithm 9 applies for the state-space model satisfying Assumption 5.5.1, provided that $\Psi_t\boldsymbol{\Sigma}_{t|t-1}\Psi_t^T + R$ are invertible matrices for $t = 1, \ldots, n$.

Exercise 5.5. Prove Proposition 5.6.1.

Exercise 5.6 (Kalman filtering of an AR(1) process observed in noise). Consider the problem studied in Exercise 4.1.

1. Can this problem be embedded in the general approach of the Kalman filter ? [*Hint:* start by comparing Equations (4.19)–(4.20) and the general state-space representation, and then discuss the assumptions and the steps carried out in Exercise 4.1.]
2. What about a noisy AR(p) model with $p \geq 2$, can the problem of its filtering be considered as a particular case of the Kalman approach ?

Chapter 6

Statistical inference

6.1 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 A 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 A.1.8 is usually referred to as *Slutsky's lemma* and is sometimes stated in the following simplest (and less general) form.

Lemma 6.1.1 (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 A.1.5 (for the weak convergence) or as in Theorem A.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 [6]. The first one indicates how to relate the weak convergence in $\mathbb{R}^p$ to the case $p = 1$.

Theorem 6.1.2 (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 6.1.3 (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,*

$$\phi_n(t) = \mathbb{E}\left[\mathrm{e}^{\mathbf{i}\mathbf{t}^T\mathbf{X}_n}\right], \quad \mathbf{t} \in \mathbb{R}^p .$$

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 6.1.4. *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 6.1.1 (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 6.1.1, 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 [6].

Theorem 6.1.5 (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 A.1.1.

In the setting of vector valued random variables, simple asymptotic results are conveniently expressed using the *stochastic order symbols*.

Definition 6.1.2 (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 A 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 A.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 6.1.6. *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 A.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 6.1).

Proposition 6.1.7. *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 6.1.8 (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. □

6.2 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 $\mathbf{X}_{1:n} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$.

To this end, we introduce two classical estimators.

Definition 6.2.1. *The* empirical mean *(or* sample mean*) and the* empirical autocovariance function *of the sample* $\mathbf{X}_{1:n}$ *are respectively defined as*

$$\widehat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t \tag{6.1}$$

$$\widehat{\Gamma}_n(h) = \begin{cases} n^{-1}\sum_{t=1}^{n-h}(\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H & \text{if } 0 \leq h \leq n-1\,, \\ n^{-1}\sum_{t=1-h}^{n}(\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H & \text{if } 0 \leq -h \leq n-1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Remark 6.2.1. *To avoid separating the different cases for* h*, the right-hand side of (6.2) can be written as follows*

$$\widehat{\Gamma}_n(h) = n^{-1}\sum_{1\leq t, t+h\leq n} (\mathbf{X}_{t+h} - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n)^H\,, \tag{6.3}$$

where, by convention, the sum is zero if there is no $t \in \mathbb{Z}$ *such that* $1 \leq t, t+h \leq n$*.*

Observe that the empirical mean (6.1) can be seen as the mean of the empirical measure, that is, denoting $\mathrm{Id}(x) = x$,

$$\widehat{\boldsymbol{\mu}}_n = P_n(\mathrm{Id})\,.$$

This is no longer true for the empirical covariance function, except in the case $h = 0$,

$$\widehat{\Gamma}_n(0) = P_n(\mathrm{Id}\,\mathrm{Id}^H) - P_n(\mathrm{Id})P_n(\mathrm{Id})^H\,.$$

Indeed the covariance relies on the distribution of the bivariate random variables $(\mathbf{X}_t, \mathbf{X}_{t+h})$, $t \in \mathbb{Z}$. Thus the empirical autocovariance function instead relies on the bivariate empirical measure, defined for all lag $h \in \mathbb{Z}$ by

$$P_{h,n}(A\times B) = n^{-1}\sum_{1\leq t, t+h\leq n} \mathbb{1}_A(\mathbf{X}_{t+h})\mathbb{1}_B(\mathbf{X}_t)\,,$$

so that

$$\widehat{\Gamma}_n(h) = P_{h,n}((\mathrm{Id} - P_n(\mathrm{Id}))\otimes(\mathrm{Id} - P_n(\mathrm{Id}))^H)\,,$$

where we used the tensor product defined by $u \otimes v(\mathbf{x}, \mathbf{y}) = u(\mathbf{x}) \times v(\mathbf{y})$. However the use of the bivariate empirical measure $P_{h,n}$ raises a difficulty because it is not a probability measure except when $h = 0$,

$$P_{h,n}(\mathbb{C}^p \times \mathbb{C}^p) = \frac{(n - |h|)_+}{n} = 1 \Leftrightarrow h = 0 \, .$$

It is thus tempting to replace the normalizing term n^{-1} by $(n - |h|)^{-1}$ in (6.2), so that we deal with empirical probability measures. However we will see that the empirical autocovariance function is a consistent estimator as $n \to \infty$ for a fixed h, in which case the two normalizations are equivalent. Moreover the normalizing term n^{-1} 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 6.2.2 (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*

$$\mathbf{I}_n(\lambda) = \frac{1}{2\pi n} \left(\sum_{t=1}^{n} (\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n) \mathrm{e}^{-\mathrm{i}t\lambda} \right) \left(\sum_{t=1}^{n} (\mathbf{X}_t - \widehat{\boldsymbol{\mu}}_n) \mathrm{e}^{-\mathrm{i}t\lambda} \right)^H \, . \tag{6.4}$$

Then we have the following result.

Theorem 6.2.1. *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 2.2.1, 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 6.2.2. *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 6.2.1 also implies that in the vector case, the empirical autocovariance function is an admissible covariance function.*

Proof of Theorem 6.2.1. 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 2.3.1, 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 2.3.3 implies that $\widehat{\Gamma}_{n,p}^{+}$ is invertible for all $p \geq 1$. □

6.3 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 6.1.5, 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 6.3.1. *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 $X_{1:n}$. 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 $\lim_{n\to\infty} \mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right] = 0$. In particular, $\widehat{\mu}_n$ is a weakly consistent estimator of μ.

(iii) If moreover $\gamma \in \ell^1$, then, as $n \to \infty$,

$$\mathrm{Var}\,(\widehat{\mu}_n) \leq n^{-1} \|\gamma\|_1 \,, \tag{6.5}$$

$$\mathrm{Var}\,(\widehat{\mu}_n) = n^{-1}(2\pi f(0) + o(1)) \,, \tag{6.6}$$

where f is the spectral density of (X_t). In particular, $\widehat{\mu}_n$ is a strongly consistent estimator of μ.

Proof. Assertion *(i)* is immediate and implies that $\mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right] = \mathrm{Var}\,(\widehat{\mu}_n)$. Thus we have

$$\begin{aligned}\mathrm{Var}\,(\widehat{\mu}_n) &= n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathrm{Cov}\,(X_s, X_t) \\ &= n^{-1}\sum_{\tau\in\mathbb{Z}}(1-|\tau|/n)_+\gamma(\tau)\;,\end{aligned}$$

where we set $\tau = s - t$ and used the notation $a_+ = \max(a, 0)$. From this expression, we easily get Assertion *(ii)* and (6.5). Under the assumption of *(iii)*, we may apply the dominated convergence and get that

$$\lim_{n\to\infty} n\,\mathrm{Var}\,(\widehat{\mu}_n) = \sum_{\tau=-\infty}^{\infty}\lim_{n\to\infty}(1-|\tau|/n)\,\gamma(\tau) = \sum_{\tau=-\infty}^{\infty}\gamma(\tau) = 2\pi f(0)\;.$$

Hence we have (6.6). Then we can show that $\widehat{\mu}_n$ is a strongly consistent estimator of μ as follows. First, (6.6) and the Markov inequality implies that, for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|\widehat{\mu}_{n^2} - \mu| \geq \epsilon\right) < \infty\;.$$

By Lemma A.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

$$\widehat{\mu}_n - \mu = \frac{m_n}{n}(\widehat{\mu}_{m_n} - \mu) + n^{-1}\sum_{s=m_n+1}^{n}(X_s - \mu)\;, \tag{6.7}$$

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 A.1.1 yields that the second term in the right-hand side of (6.7) 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 6.3.1 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 6.3.1. *$X = (X_t)_{t\in\mathbb{Z}}$ is a real valued* linear process with short memory*, that is, it admits the representation*

$$X = \mu + \mathrm{F}_\psi(Z) , \tag{6.8}$$

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

$$\gamma(h) = \sigma^2 \sum_{k\in\mathbb{Z}} \psi_{k+h}\psi_k \tag{6.9}$$

$$f(\lambda) = \frac{1}{2\pi}\sum_{\tau\in\mathbb{Z}} \gamma(\tau)\mathrm{e}^{-\mathrm{i}\tau\lambda} . \tag{6.10}$$

Now, to compute the 4th order moments of X, we shall just need an assumption on Z. We shall use the following one.

Assumption 6.3.2. *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 6.3.2 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 6.2).

Lemma 6.3.2. *Suppose that Assumption 6.3.1 and Assumption 6.3.2 hold with $\mu = 0$. Then, for all $k, l, p, q \in \mathbb{Z}$*

$$\begin{aligned}\mathbb{E}\left[X_k X_\ell X_p X_q\right] = (\eta-3)\sigma^4 \sum_{i\in\mathbb{Z}} \psi_{k+i}\psi_{\ell+i}\psi_{p+i}\psi_{q+i} + \gamma(k-\ell)\gamma(p-q) \\ + \gamma(k-p)\gamma(\ell-q) + \gamma(k-q)\gamma(\ell-p)\ , \end{aligned} \tag{6.11}$$

where γ is the autocovariance function of X. Moreover, there exists a constant C such that, for all $m \in \mathbb{N}$,

$$\mathbb{E}\left[\left(\sum_{t=1}^{m} X_t\right)^4\right] \leq Cm^2\ . \tag{6.12}$$

Let us now state a weak consistency result for the empirical covariance function of a real valued process.

Theorem 6.3.3. *Suppose that Assumption 6.3.1 and Assumption 6.3.2 hold. Let $\widehat{\gamma}_n$ denote the empirical autocovariance function of the sample $X_{1:n}$. Then, for all $p, q \in \mathbb{Z}$,*

$$\mathbb{E}\left[\widehat{\gamma}_n(p)\right] = \gamma(p) + O(n^{-1})\ , \tag{6.13}$$

$$\lim_{n\to\infty} n\,\mathrm{Cov}\left(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q)\right) = V(p,q)\ , \tag{6.14}$$

where γ is the autocovariance function of X and

$$V(p,q) = (\eta-3)\gamma(p)\gamma(q) + \sum_{u\in\mathbb{Z}} \left[\gamma(u)\gamma(u-p+q) + \gamma(u+q)\gamma(u-p)\right]\ . \tag{6.15}$$

In particular, $\widehat{\gamma}_n(p)$ is a weakly consistent estimator of $\gamma(p)$,

$$\widehat{\gamma}_n(p) = \gamma(p) + O_P(n^{-1/2})\ . \tag{6.16}$$

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, (6.13) and (6.14) yield (6.16). Hence it only remains to prove (6.13) and (6.14).

To this end, we introduce

$$\tilde{\gamma}_n(h) = n^{-1} \sum_{t=1}^{n} X_{t+h} X_t \ . \tag{6.17}$$

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 6.3.2, 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 $h \in \mathbb{Z}$,

$$\mathbb{E}\left[\left(\widehat{\gamma}_n(h) - \tilde{\gamma}_n(h) \right)^2 \right] = O(n^{-2}) \tag{6.18}$$

By Jensen's inequality, we get

$$\mathbb{E}\left[\widehat{\gamma}_n(p)\right] = \mathbb{E}\left[\tilde{\gamma}_n(p)\right] + O(n^{-1}) \ , \tag{6.19}$$

Since $\tilde{\gamma}_n(p)$ is an unbiased estimator of $\gamma(p)$, this yields (6.13). Next, we have, for all $p, q \in \mathbb{Z}$,

$$\mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) = n^{-2} \sum_{s=1}^{n} \sum_{t=1}^{n} \mathrm{Cov}\left(X_{s+p} X_s, X_{t+q} X_t\right) \ .$$

By Lemma 6.3.2, 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

$$v(u) = (\eta - 3)\sigma^4 \sum_{i \in \mathbb{Z}} \psi_{u+i} \psi_{u+p+i} \psi_i \psi_{q+i} + \gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q) \ . \tag{6.20}$$

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 (6.9) and (6.15), we have

$$\sum_{\tau \in \mathbb{Z}} v(\tau) = V(p, q) \; .$$

Hence we get that, as $n \to \infty$,

$$\mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) \sim n^{-1} V(p, q) \; . \tag{6.21}$$

From this and (6.18), it follows that, for all $p, q \in \mathbb{Z}$,

$$\mathrm{Cov}\left(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q)\right) = \mathrm{Cov}\left(\tilde{\gamma}_n(p), \tilde{\gamma}_n(q)\right) + O(n^{-3/2}) \; . \tag{6.22}$$

Finally, (6.21) and (6.22) yield (6.14), which concludes the proof. □

6.4 Asymptotic distribution of the empirical mean

From Theorem 6.3.1, 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 6.1.5) 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 6.3.3, we need to somehow precise the distribution of the time series by relying on Assumption 6.3.1 with a suitable assumption on the white noise Z, namely.

Assumption 6.4.1. *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 6.1.5 is given by the following result.

Proposition 6.4.1. *Suppose that Assumption 6.3.1 and Assumption 6.4.1 hold with a finitely supported sequence ψ. Let $\widehat{\mu}_n$ denote the empirical mean of the sample $X_{1:n}$. Then, as $n \to \infty$,*

$$\sqrt{n}\,(\widehat{\mu}_n - \mu) \Longrightarrow \mathcal{N}(0, 2\pi f(0)) \tag{6.23}$$

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 (6.23). □

Now, we can state a more general extension similar to Proposition 6.4.1 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 6.4.2. *Suppose that Assumption 6.3.1 and Assumption 6.4.1 hold. Let $\widehat{\mu}_n$ denote the empirical mean of the sample $X_{1:n}$. Then the CLT (6.23) holds.*

Proof. As in the proof of Proposition 6.4.1, we can assume that $\mu = 0$ without loss of generality.

Define the sequence ψ^m by

$$\psi_k^m = \begin{cases} \psi_k & \text{if } |k| \le m, \\ 0 & \text{otherwise.} \end{cases} \tag{6.24}$$

Let $\widehat{\mu}_n^m$ be the empirical mean of the the sample $[\mathrm{F}_{\psi^m}(Z)]_{1:n}$. Then by Proposition 6.4.1, we have, for all $m \ge 1$, as $n \to \infty$,

$$\sqrt{n}\,(\widehat{\mu}_n^m - \mu) \Longrightarrow \mathcal{N}(0, \sigma_m^2)\ , \tag{6.25}$$

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 6.1.4, as $n \to \infty$,

$$\mathcal{N}(0, \sigma_m^2) \Longrightarrow \mathcal{N}(0, 2\pi f(0)) .$$

By Lemma A.1.9, this convergence and (6.25) imply (6.23) if for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}|\widehat{\mu}_n - \widehat{\mu}_n^m| > \epsilon) = 0 . \tag{6.26}$$

Hence it only remains to show (6.26). 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 6.3.1, 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 (6.26) follows by the Markov inequality. □

Proposition 6.4.1 and Theorem 6.4.2 heavily rely on the linear representation of Assumption 6.3.1. It is interesting to note that the above technique can be applied in the following framework which do not assume a linear representation.

Definition 6.4.1. *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 6.4.3. *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 (6.23) 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 A.1.9, as in the proof of Theorem 6.4.2. 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 6.1.5 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 A.1.9, to conclude the proof, it only remains to show that, for all $\epsilon > 0$,

$$\lim_{p\to\infty} \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}|R_{n,p}| > \epsilon) = 0 \, . \tag{6.27}$$

We first observe that, since $r \leq p + m$ for all n, we have

$$\mathrm{Var}\left(\sum_{s=1}^{r} X_{k(p+m)+s} \right) \leq C \, (p + m) \, , \tag{6.28}$$

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 (6.28) 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 (6.27) and the proof is finished. □

6.5 Asymptotic distribution of the empirical autocovariance function

An approach similar to Theorem 6.4.2 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 (6.8).

Assumption 6.5.1. *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 6.3.2. Thus the asymptotic behavior of the covariances $\mathrm{Cov}\,(\widehat{\gamma}_n(p), \widehat{\gamma}_n(q))$ are given by Theorem 6.3.3. 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 (6.14). This result is stated in the following theorem.

Theorem 6.5.1. *Suppose that Assumption 6.3.1 and Assumption 6.5.1 hold. Let $\widehat{\gamma}_n$ denote the empirical autocovariance function of the sample $X_{1:n}$. Then, as $n \to \infty$,*

$$\sqrt{n}\ (\widehat{\gamma}_n - \gamma) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V)\ , \tag{6.29}$$

where V is defined by (6.15). As a consequence, we also have, as $n \to \infty$,

$$\sqrt{n}\ (\widehat{\rho}_n - \rho) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, W)\ , \tag{6.30}$$

where $\widehat{\rho}_n(h) = \widehat{\gamma}_n(h)/\widehat{\gamma}_n(0)$ and $\rho(h) = \gamma(h)/\gamma(0)$.

$$W(p,q) = \sum_{u=1}^{\infty} \{\rho(u+p) + \rho(u-p) - 2\rho(u)\rho(p)\} \times \{\rho(u+q) + \rho(u-q) - 2\rho(u)\rho(q)\}\ . \tag{6.31}$$

Proof. The CLT (6.30) follows from (6.29) (see Exercise 6.3), so we only show (6.29).

As in the proof of Theorem 6.3.3, we can take $\mu = 0$ without loss of generality. We also observe that (6.18) implies that $\widehat{\gamma}_n = \tilde{\gamma}_n + O_p(n^{-1})$. Hence it is sufficient to prove that, as $n \to \infty$,

$$\sqrt{n}\ (\tilde{\gamma}_n - \gamma) \overset{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V)\ , \tag{6.32}$$

The proof of this is left to the reader, see Exercise 6.4. □

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*.

6.6 Application to ARMA processes

Let us give some applications of the above asymptotic results on some examples of ARMA processes.

Example 6.6.1 (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 6.4.2 and Theorem 6.5.1 apply. Note that $\rho(h) = 0$ for all $h \neq 0$ and $W(p,q) = \mathbb{1}_{\{p=q\}}$. Hence (6.30) 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[6]. 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 6.6.2 (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 6.4.2 and Theorem 6.5.1 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 6.6.3 (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 6.4.2 are satisfied and the limit variance in (6.23) 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 6.5.1 also hold. A direct computation yields

$$\begin{aligned} W(h,h) &= \sum_{m=1}^{h} \phi^{2h}(\phi^{-m} - \phi^m)^2 + \sum_{m=h+1}^{\infty} \phi^{2m}(\phi^{-i} - \phi^i)^2 \\ &= (1-\phi^{2h})(1+\phi^2)(1-\phi^2)^{-1} - 2h\phi^{2h} \end{aligned}$$

6.7 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 6.7.1 (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

$$\widehat{\theta}_n \overset{\mathrm{def}}{=} \operatorname*{argmin}_{\theta\in\Theta} - \log L_n(\theta) \, , \tag{6.33}$$

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

$$-\log L_n(\widehat{\theta}_n) \leq \inf_{\theta\in\Theta} -L_n(\theta) + o_P(n^{-1/2}) \ . \tag{6.34}$$

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

$$-\log L_n(\theta) = -\sum_{t=1}^{n} \log p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}, \theta) \ . \tag{6.35}$$

Under the *usual regular assumption* (see [1]), the Information matrix is defined as

$$I_n(\theta) = \mathrm{Cov}\left(\partial \log L_n(\theta)|\theta\right) = -\mathbb{E}\left[\partial\partial^T \log L_n(\theta)|\theta\right] \ , \tag{6.36}$$

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 (6.35) and (6.36), $I_n(\theta)$ 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 (6.33) (or satisfying (6.34)) 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,

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \Longrightarrow \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta)\right) \tag{6.37}$$

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, 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

$$\boldsymbol{\eta}_t(\theta) = \mathbf{Z}_t - \mathbb{E}\left[\mathbf{Z}_t \middle| \mathbf{Z}_{1:t-1}, \theta\right] , \tag{6.38}$$

$$\widetilde{\Sigma}_t(\theta) = \mathrm{Cov}\left(\mathbf{Z}_t - \boldsymbol{\eta}_t(\theta) | \theta\right) , \tag{6.39}$$

Hence (6.35) yields

$$-2\log L_n(\theta) = n\log(2\pi) + \sum_{t=1}^{n} \log\det\widetilde{\Sigma}_t(\theta) + \sum_{t=1}^{n} \boldsymbol{\eta}_t(\theta)^T \widetilde{\Sigma}_t(\theta)^{-1} \boldsymbol{\eta}_t(\theta) . \tag{6.40}$$

Denoting by ∂_i the derivative with respect to the i-th component of θ, the gradient is then given by

$$-2\partial_i \log L_n = \sum_{t=1}^{n} \left\{ \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\right) + 2\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\boldsymbol{\eta}_t] - \boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \right\}, \tag{6.41}$$

and the Hessian matrix by

$$\begin{aligned} -2\partial_i\partial_j \log L_n = \sum_{t=1}^{n} \Big\{ & \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i\partial_j\widetilde{\Sigma}_t]\right) - \mathrm{Trace}\left(\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\right) \\ & + 2\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\partial_j\boldsymbol{\eta}_t] + 2[\partial_i\boldsymbol{\eta}_t^T]\widetilde{\Sigma}_t^{-1}[\partial_j\boldsymbol{\eta}_t] - 2\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\boldsymbol{\eta}_t] \\ & - 2\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_i\boldsymbol{\eta}_t] - \boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\partial_j\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \\ & + 2\boldsymbol{\eta}_t^T\widetilde{\Sigma}_t^{-1}[\partial_i\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}[\partial_j\widetilde{\Sigma}_t]\widetilde{\Sigma}_t^{-1}\boldsymbol{\eta}_t \Big\} . \end{aligned} \tag{6.42}$$

Observe that by (6.38), 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 (6.42), 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

$$I_n(i,j;\theta) = \sum_{t=1}^{n} \left\{ \mathbb{E}\left[\left\{ [\partial_i \boldsymbol{\eta}_t^T] \widetilde{\Sigma}_t^{-1} [\partial_j \boldsymbol{\eta}_t] \right\}(\theta) \middle| \theta \right] + \frac{1}{2} \mathrm{Trace}\left(\left\{ \widetilde{\Sigma}_t^{-1} [\partial_i \widetilde{\Sigma}_t] \widetilde{\Sigma}_t^{-1} [\partial_j \widetilde{\Sigma}_t] \right\}(\theta) \right) \right\} . \quad (6.43)$$

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 (6.37), 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 $\widehat{I}_n$).

6.8 Exercises

Exercise 6.1. 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 6.2. Let $(X_t)_{t\in\mathbb{Z}}$ satisfy Assumption 6.3.1 and Assumption 6.3.2 with $\mu = 0$.

1. Show that (6.11) 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 (6.12) holds.

Exercise 6.3. Use the δ-method to show that (6.29) implies (6.30).

Exercise 6.4. Suppose that Assumption 6.3.1 and Assumption 6.5.1 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 (6.17). For any $m \geq 1$, define $X^{(m)} = \mathrm{F}_{\psi^m}(Z)$ with ψ^m defined by (6.24) and $\tilde{\gamma}_n^{(m)}$ be defined as in (6.17) 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 6.4.3 to show that

$$\sqrt{n}\ \left(\tilde{\gamma}_n^{(m)} - \gamma_m\right) \stackrel{\text{fidi}}{\Longrightarrow} \mathcal{N}(0, V_m) \ ,$$

   where V_m is defined by (6.15).

3. Proceed as in the proof of Theorem 6.3.3 to show that

$$\lim_{m\to\infty} \limsup_{n\to\infty} n\ \text{Var}\left(\tilde{\gamma}_n - \tilde{\gamma}_n^{(m)}\right) = 0 \ .$$

4. Use Lemma A.1.9 to conclude the proof of Theorem 6.5.1.

Exercise 6.5. Let (X_t) be a weakly stationary real valued process with mean μ and autocovariance function γ. We observe $X_1, \ldots, X_n$.

1. Determine the linear unbiased estimator $\hat{\mu}_n$ of μ that minimizes the risk

$$\text{EQM} = \mathbb{E}[(\mu - \hat{\mu}_n)^2].$$

2. Give the corresponding risk.

Exercise 6.6 (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 $t \geq 1$,

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

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 $X_1, \ldots, X_n$.

1. Show that empirical covariance matrices $\hat{\Gamma}_t$ are invertible for all $t \geq 1$ under a simple condition on $X_1, \ldots, X_n$. [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 $(\varepsilon_t)_{t\in\mathbb{Z}} \sim \text{WN}(0, \sigma^2)$. Suppose that we observe a sample $X_1, \ldots, X_n$ 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 6.7 (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:

$$\begin{cases} I_{1,\theta} = X_1, \quad \upsilon_{1,\theta} = (1+\rho^2)\sigma^2 \\ I_{t,\theta} = X_t - \rho\frac{\sigma^2}{\upsilon_{t-1,\theta}} I_{t-1,\theta}, \quad \upsilon_{t,\theta} = (1+\rho^2)\sigma^2 - \frac{\rho^2\sigma^4}{\upsilon_{t-1,\theta}} \quad \text{for } t = 2, \ldots, n \end{cases}$$

7. Considering $\tilde{\upsilon}_{t,\theta} = \upsilon_{t,\theta}/\sigma^2$, obtain the expression of $\hat{\sigma}_n^2$ as a function of $\hat{\rho}_n$ and of the observations $X_1, \ldots, X_n$.

8. Show that, for all $\theta \in \Theta$, $\tilde{\upsilon}_{t,\theta} \to 1$.

Appendix A

Convergence of random variables 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 [2]. Here we provide a brief account of the essential classical definitions and results. The detailed proofs can be found in [2].

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 $\mathcal{B}(\mathsf{X})$.

A.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 A.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 A.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 A.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 A.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

$$\mu(A) = \inf\{\mu(U) \; : \; U \text{ open set } \supset A\} = \sup\{\mu(F) \; : \; F \text{ closed set } \subset A\} \; . \tag{A.1}$$

Proposition A.1.3. *Let $\mu \in \mathbb{M}_1(\mathsf{X})$. Then (A.1) holds for all $A \in \mathcal{B}(\mathsf{X})$.*

Definition A.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 A.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 A.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

$$D_h \stackrel{\text{def}}{=} \{x \in \mathsf{X} : \ h \text{ is discontinuous at } x\} \ . \tag{A.2}$$

The following theorem is often referred to as the *continuous mapping theorem.*

Theorem A.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 A.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 A.1.5 and Theorem A.1.2.

Theorem A.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 (A.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 A.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 A.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

$$\begin{aligned} \mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right] &= \mathbb{E}\left[f(X_n) - f(Z_{m,n})\right] \\ &\quad + \left[\mathbb{E}\left[f(Z_{m,n})\right] - \mathbb{E}\left[f(Z_m)\right]\right] + \left[\mathbb{E}\left[f(Z_m)\right] - \mathbb{E}\left[f(Z)\right]\right] . \quad \text{(A.3)} \end{aligned}$$

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 A.1.4 and using the assumptions of the lemma, we get that, for some large enough m,

$$\limsup_{n\to\infty} |\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(Z)\right]| \leq (K\,\epsilon + C)\epsilon + 0 + \epsilon\,.$$

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 A.1.4. □

A.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 $\lambda, \mu \in \mathbb{M}_1(\mathsf{X})$,

$$\boldsymbol{\rho}(\lambda, \mu) = \inf\left\{\alpha > 0 \,:\, \lambda(F) \leq \mu(F^\alpha) + \alpha \text{ for all } F \in \mathcal{S}\right\}. \tag{A.4}$$

The following result shows that $\boldsymbol{\rho}$ is indeed a metric, which is not completely obvious from (A.4).

Lemma A.2.1. *$\boldsymbol{\rho}$ defined in (A.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 A.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 A.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 A.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 A.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 A.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 A.2.5. *Let* (X, d) *be separable and complete. Then* $(\mathbb{M}_1(\mathsf{X}), \boldsymbol{\rho})$ *is separable and complete.*

Index

Bibliography

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York, NY, third edition, 2003.

[2] P. Billingsley. *Convergence of probability measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[3] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, 2nd edition, 1991.

[4] P. E. Caines. *Linear Stochastic Systems.* Wiley, 1988.

[5] E. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* John Wiley & Sons, 1988.

[6] J. Jacod and P. Protter. *Probability essentials.* Universitext. Springer-Verlag, Berlin, second edition, 2003.

[7] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[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.

[9] R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications.* New York: Springer, 3rd edition, 2011.

[10] The R project for statistical computing. http://www.r-project.org/.

[11] N. Young. *An introduction to Hilbert space.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1988.