Institut Mines-Télécom

# Linear time series: Spectral analysis

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr



TELECOM Paris
IP PARIS

# Contents



**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# List of Figures

# Acronyms

**ACF** *Auto-Covariance Function*

**AIC** *Akaike Information Criterion*

**AR** *Autoregressive*

**ARMA** *Autoregressive Moving Average*

**CRLB** *Cramér-Rao Lower Bound*

**dB** *Decibels*

**DFT** *Discrete Fourier Transform*

**DTFT** *Discrete Time Fourier Transform*

**EDC** *Efficient Detection Criteria*

**EDS** *Exponentially Damped Sinusoids*

**ESM** *Exponential Sinusoidal Model*

**ESPRIT** *Estimation of Signal Parameters via Rotational Invariance Techniques*

**EVD** *EigenValue Decomposition*

**FFT** *Fast Fourier Transform*

**FIR** *Finite Impulse Response*

**HR** *High Resolution*

**ITC** *Information Theoretic Criteria*

**LS** *Least Squares*

**MA** *Moving Average*

**MDL** *Minimum Description Length*

**MUSIC** *MUltiple SIgnal Classification*

**MVDR** *Minimum Variance Distortionless Response*

**PSD** *Power Spectral Density*

**SNR** *Signal to Noise Ratio*

**SVD** *Singular Value Decomposition*

**WN** *White Noise*

**WSS** *Wide Sense Stationary*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# Mathematical notation

$\mathbb{N}$ set of natural numbers

$\mathbb{Z}$ set of integers

$\mathbb{R}$ set of real numbers

$\mathbb{C}$ set of complex numbers

$\triangleq$ equal by definition to

$\mathcal{R}e(.)$ real part

$\mathcal{I}m(.)$ imaginary part

$x$ (normal font, lower case) scalar

$\boldsymbol{x}$ (bold font, lower case) vector

$\boldsymbol{A}$ (bold font, upper case) matrix

diag($\boldsymbol{x}$) diagonal matrix whose diagonal entries are those of vector $\boldsymbol{x}$

$.^{\top}$ transpose of a matrix

$\overline{(.)}$ conjugate of a matrix / vector / number

$.^{H}$ conjugate transpose of a matrix

span(.) range space of a matrix

ker(.) kernel of a matrix

dim(.) dimension of a vector space

rank(.) rank of a matrix

trace(.) trace of a square matrix

det(.) determinant of a square matrix

$\mathbf{1}_I(x)$ indicator function of set $I$ ($\mathbf{1}_I(x) = 1$ if $x \in I$, and $\mathbf{1}_I(x) = 0$ otherwise)

$*$ convolution product between two sequences

$\widetilde{.}$ conjugation and time reversal of time series ($\widetilde{h}_t = \overline{h_{-t}}$)

$\mathbb{E}[.]$ mathematical expectation

var[.] variance

$\widehat{(.)}$ estimator of a parameter

CRB{.} Cramér-Rao bound

WN$(0, \sigma^2)$ centered *White Noise* (WN) of variance $\sigma^2$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# Chapter 1

# Introduction

Spectral analysis considers the problem of determining the spectral content (i.e., the distribution of power over frequency) of a time series from a finite set of measurements, by means of either non-parametric or parametric techniques. The history of spectral analysis as an established discipline started more than a century ago with the work by Schuster on detecting cyclic behavior in time series. An interesting historical perspective on the developments in this field can be found in [1]. This reference notes that the word "spectrum" was apparently introduced by Newton in relation to his studies of the decomposition of white light into a band of light colors, when passed through a glass prism. This word appears to be a variant of the Latin word "specter" which means "ghostly apparition".

This course is the continuation of the teaching unit TSIA202a. From time to time, we will thus present reminders from TSIA202a. Their proof will be omitted, since they can be found in the course handout of TSIA202a. Such reminders will be highlighted by a bar on the left side of the page, as follows:

| This is a reminder from TSIA202a.

Besides, this course was partly elaborated from a few books, mostly [2]. In some places, we will thus refer to these books, for those of you who wish to go deeper into the course. These references will be highlighted by a curved arrow on the left side of the page, as follows:

↪ For more details on the content of this course, we refer the reader to [2].

## 1 Parametric and non-parametric spectral analysis

The essence of the spectral estimation problem is captured by the following informal formulation: *From a finite record of a stationary data sequence, estimate how the total power is distributed over frequency.* Spectral analysis finds applications in many diverse fields. In vibration monitoring, the spectral content of measured signals give information on the wear and other characteristics of mechanical parts under study. In economics, meteorology, astronomy and several other fields, the spectral analysis may reveal "hidden periodicities" in the studied data, which are to be associated with cyclic behavior or recurring processes. In speech analysis, spectral models of voice signals are useful in better understanding the speech production process, and – in addition – can be used for both speech synthesis (or compression) and speech recognition. In radar and sonar systems, the spectral contents of the received signals provide information on the location of the sources (or targets) situated in the field of view. In medicine, spectral analysis of various signals measured from a patient, such as electrocardiogram (ECG) or electroencephalogram (EEG) signals, can provide useful material for diagnosis. In seismology, the spectral analysis of the signals recorded prior to and during a seismic event (such as a volcano eruption or an earthquake) gives useful information on the ground movement associated with such events and may help in predicting them. Seismic spectral estimation is also used to predict sub-surface geologic structure in gas and oil exploration. In control systems, there is a resurging interest in spectral analysis methods as a means of characterizing the dynamical behavior of a given system, and ultimately synthesizing a controller for that system. The textbook [1] contains a well-written historical perspective on spectral estimation which is worth reading.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

There are two broad approaches to spectral analysis. One of these derives its basic idea directly from the above definition: the studied signal is applied to a band-pass filter with a narrow bandwidth, which is swept through the frequency band of interest, and the filter output power divided by the filter bandwidth is used as a measure of the spectral content of the input to the filter. This is essentially what the classical (or non-parametric) methods of spectral analysis do. These methods are described in Chapters 2 and 4 of this document. The second approach to spectral estimation, called the parametric approach, is to postulate a model for the data, which provides a means of parameterizing the spectrum, and to thereby reduce the spectral estimation problem to that of estimating the parameters in the assumed model. The parametric approach to spectral analysis is treated in Chapters 3 and 5. Parametric methods may offer more accurate spectral estimates than the non-parametric ones in the cases where the data indeed satisfy the model assumed by the former methods. However, in the more likely case that the data do not satisfy the assumed models, the non-parametric methods may outperform the parametric ones owing to the sensitivity of the latter to model misspecifications. This observation has motivated renewed interest in the non-parametric approach to spectral estimation.

Many real-world signals can be characterized as being random (from the observer's viewpoint). Briefly speaking, this means that the variation of such a signal outside the observed interval cannot be determined exactly but only specified in statistical terms of averages. In this text, we will be concerned with estimating the spectral characteristics of random signals. Throughout this work, we consider discrete signals (or data sequences). Such signals are most commonly obtained by the temporal or spatial sampling of a continuous (in time or space) signal.

# 2 Line spectrum estimation

Among parametric spectral analysis methods, the ones dedicated to line spectrum estimation stand apart. In the context of speech and music signal processing, the tonal part of a wide variety of sounds is accurately modeled as a sum of sinusoids with slowly varying parameters. For example, the sounds that produce a well-defined perception of pitch have a quasi-periodic waveform (over a duration greater than a few tens of miliseconds). Fourier analysis shows that these signals are composed of sinusoids satisfying a relation of *harmonicity*, which means that their frequencies are multiples of the fundamental frequency, defined as the inverse of the period. This is the case of voiced speech signals produced by the quasi-periodic vibration of the vocal cords, such as vowels. Many wind or string instruments also produce harmonic or quasi-harmonic sounds. However, in a polyphonic music signal, the sounds emitted simultaneously by one or more instruments overlap; thus the harmonic relationship is no longer verified, but the signal remains essentially made up of sinusoids.

The estimation of sinusoids is a classic problem, more than two hundred years old. In this area, the Fourier transform is a privileged tool because of its robustness, the simplicity of its implementation, and the existence of fast algorithms (*Fast Fourier Transform* (FFT)). However, it has a number of drawbacks. First, its frequency *precision*, that is, the precision with which the frequency of a sinusoid can be estimated, is limited by the number of samples used to calculate it. This first limitation can be circumvented by extending the signal by a series of zeros (this operation is called *zero-padding*). However, its frequency *resolution*, that is to say its ability to distinguish two close sinusoids, remains limited by the duration of the observed signal. Despite these drawbacks, the Fourier transform remains the most used tool in spectral analysis. It has given rise to numerous frequency estimation methods [3].

The *High Resolution* (HR) methods, which find their applications both in antenna processing and in spectral analysis [4], have the advantage of overcoming the natural limitations of Fourier analysis. Indeed, in the absence of noise, their frequency precision and resolution are virtually infinite (although in practice limited by the finite precision of computers). This is made possible by exploiting a parametric signal model. Thus, unlike the Fourier analysis which consists in representing the signal in a transformed domain, the HR methods are parametric estimation methods. In the context of audio signal processing, despite their superiority in terms of spectral resolution (in particular over short time windows), they remain little used because of their high computational complexity. Nevertheless, the HR methods are well suited for estimating the parameters of a sum of sinusoids whose amplitudes vary exponentially (*Exponential Sinusoidal Model* (ESM) model). This type of modulation makes it possible to describe the natural damping of free vibratory systems, such as the vibration of a plucked string [5]. On the other hand, it has been shown in [6] that the HR methods prove to be particularly effective in the case of strongly

attenuated signals. More generally, the ESM model makes it possible to describe signals with large amplitude variations [7]. In addition, the music signals often contain pairs or triplets of very close frequencies which generate a beat phenomenon. These beats strongly contribute to the natural appearance of the sound. They often result from the special properties of vibration systems. For example, a minor asymmetry in the geometry of a bell leads to pairs of vibration modes. In the case of a guitar, the coupling between the strings and the bridge can be represented by a so-called mobility matrix, from which it is possible to deduce frequency pairs [8]. In the case of the piano, the coupling of the horizontal and vertical vibration modes of each string and the presence of pairs or triplets of strings for most notes explain the presence of four or six neighboring frequencies at the level of each harmonic [9]. The Fourier analysis generally does not make it possible to distinguish all these frequencies. Studies in [6] on piano and guitar sounds have shown the superiority of HR methods in this area. The same technique was used to estimate physical parameters, such as the radiation factor of a guitar [10], and to study the propagation of mechanical waves in solid materials [11].

# 3 Kalman filter

The Kalman filter is an infinite impulse response filter that estimates the states of a dynamic system from a series of incomplete or noisy measurements. The filter was named after the Hungarian-born American mathematician and computer scientist Rudolf Kalman.

The Kalman filter is used in a wide range of technological fields (radar, electronic vision, communication...). It is a major theme in automatic control and signal processing. An example of use can be the continuous providing of information such as the position or the speed of an object from a series of observations relative to its position, possibly including measurement errors. For example, in the case of radars where we want to track a target, data on its position, velocity and acceleration are measured at each moment but with a lot of perturbations due to noise or measurement errors. The Kalman filter uses the dynamics of the target which defines its evolution in time to obtain better data, thus eliminating the effect of noise. These data can be computed in the present (filtering), in the past (smoothing), or on a future horizon (prediction).

Kalman filtering is also increasingly used outside the field of electronics, for example in meteorology and oceanography, for data assimilation in a numerical model, in finance or in navigation, and it is even used in the estimation of road traffic conditions in the case of ramp control where the number of magnetic loops on the road is insufficient.

# 4 Outline

This document is organized as follows. Chapter 2 introduces the spectral analysis problem, motivates the definition of power spectral density functions, and reviews some important properties of autocovariance sequences and spectral density functions. Then it addresses non-parametric spectral estimation and presents classical techniques, including the periodogram, the correlogram, and their modified versions to reduce variance. We include an analysis of bias and variance of these techniques, and relate them to one another. Chapter 3 considers parametric techniques; it focuses on continuous spectral models (Autoregressive Moving Average (ARMA) models and their AR and MA special cases). Chapter 4 considers the more recent filter bank version of non-parametric techniques, including both data-independent and data-dependent filter design techniques. The Capon estimator introduced in Chapter 4 has interpretations as both an averaged AR spectral estimator and as a matched filter for line spectral models, that will be addressed in Chapter 5, which focuses on discrete spectral models (sinusoids in noise). Finally, the Kalman filter is presented in Chapter 6.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# Chapter 2

# Non-parametric spectral estimation

The non-parametric methods of spectral estimation constitute the "classical means" for *Power Spectral Density* (PSD) estimation of *Wide Sense Stationary* (WSS) processes. The present chapter reviews the main non-parametric methods and their properties. A related discussion is to be found in Chapter 4, where the non-parametric approach to PSD estimation is given a filter bank interpretation. We first introduce two common spectral estimators, the periodogram and the correlogram. These methods are then shown to be equivalent under weak conditions. The periodogram and correlogram methods provide reasonably high resolution for sufficiently long data lengths, but are poor spectral estimators because their variance is high and does not decrease with increasing data length. The high variance of the periodogram and correlogram methods motivates the development of modified methods that have lower variance, at a cost of reduced resolution. Several modified methods have been introduced, and we present some of the most popular ones. We show them all to be more-or-less equivalent in their properties and performance for large data lengths.

The chapter is structured as follows: we will first present a few useful reminders about WSS processes, and the estimation of their mean and their *Auto-Covariance Function* (ACF) in Section 1. Section 2 will then be devoted to PSD estimators, namely the periodogram and its variants (Blackman-Tukey, Bartlett, Welch and Daniell), and their performance analysis in terms of bias and variance.

↪ For more details on the content of this chapter, we refer the reader to [2, chap. 2].

## 1 Reminder: WSS processes and statistical inference

The results presented in this section are reminders of the teaching unit TSIA 202a.

### 1.1 Wide Sense Stationary (WSS) processes

**Definition 1** (WSS process)**.** *A complex WSS process is a sequence of random variables $X_t \in \mathbb{C}$ for $t \in \mathbb{Z}$, such that*

- $\mathbb{E}[|X_t|^2] < +\infty$*;*
- $\mathbb{E}[X_t] = \mu_X$ *does not depend on $t$;*
- $\forall k \in \mathbb{Z}$, $\mathrm{cov}[X_{t+k}, X_t] = \mathbb{E}[X^c_{t+k}\overline{X^c_t}]$ *(where $X^c_t = X_t - \mu_X$) does not depend on $t$;.*

**Proposition 1** (Strict vs. wide-sense stationarity)**.** *We have the following relationships between wide sense and strict stationarities:*

- *if $X_t$ is a strictly stationary process such that $\mathbb{E}[|X_t|^2] < +\infty$, then $X_t$ is also WSS;*
- *if $X_t$ is a Gaussian process, then it is strictly stationary if and only if it is WSS.*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Contexte académique } sans modifications
*Voir Page 98*

**Definition 2** (Autocovariance function)**.** *Let $X_t$ be a complex WSS process. Its ACF is defined as*

$$\forall k \in \mathbb{Z},\ r_{XX}(k) = \mathrm{cov}[X_{t+k}, X_t] = \mathbb{E}[X^c_{t+k}\overline{X^c_t}].$$

**Proposition 2** (Properties of the autocovariance function)**.** *Let $X_t$ be a complex WSS process. Then its ACF is such that*

- $r_{XX}(0) = \mathrm{var}[X_t] \geq 0$*;*
- $\forall k \in \mathbb{Z}$, $r_{XX}(-k) = \overline{r_{XX}(k)}$ *(Hermitian symmetry);*
- $\forall k \in \mathbb{N}^*$, $\forall t_1 \dots t_k \in \mathbb{Z}$, $\forall \lambda_1 \dots \lambda_k \in \mathbb{C}$, $\sum_{i=1}^{k}\sum_{j=1}^{k} \lambda_i\overline{\lambda_j} r_{XX}(t_i - t_j) \geq 0$ *(positive semi-definiteness);*
- $\forall k \in \mathbb{Z}$, $|r_{XX}(k)| \leq r_{XX}(0)$ *(boundedness);*
- $P_X \triangleq \mathbb{E}[|X_t|^2] = r_{XX}(0) + |\mu_X|^2$ *(power of a WSS process).*

**Definition 3** (Power spectral density)**.** *Let $X_t$ be a complex WSS process. If $r_{XX} \in l^1(\mathbb{Z})$, then the PSD of $X_t$ is defined as the* Discrete Time Fourier Transform *(DTFT) of $r_{XX}$:*

$$\forall \nu \in \mathbb{R},\ S_{XX}(\nu) = \sum_{k=-\infty}^{+\infty} r_{XX}(k) e^{-2\imath\pi\nu k}.$$

**Proposition 3** (Properties of the power spectral density)**.** *Let $X_t$ be a complex WSS process such that $r_{XX} \in l^1(\mathbb{Z})$. Then:*

- $r_{XX}(k) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} S_{XX}(\nu) e^{+2\imath\pi\nu k} d\nu$ *(inversion of the DTFT);*
- $\nu \mapsto S_{XX}(\nu)$ *is a continuous function (as the DTFT of a $l^1(\mathbb{Z})$ sequence);*
- $S_{XX}(\nu) \geq 0\ \forall \nu \in \mathbb{R}$ *(due to the Herglotz theorem);*
- $P_X = r_{XX}(0) + |\mu_X|^2 = \int_{-\frac{1}{2}}^{+\frac{1}{2}} S_{XX}(\nu) d\nu + |\mu_X|^2$ *(power of a WSS process).*

**Theorem 4** (Filtering theorem for WSS processes)**.** *Let $h_k$ be the impulse response of a stable filter of frequency response $H(\nu) = \sum_{k=-\infty}^{+\infty} h_k e^{-2\imath\pi\nu k}$, let $X_t$ be a complex WSS process, and let $Y_t = h * X_t$. Then $Y_t$ is a WSS process of mean $\mu_Y = \mu_X H(0) = \mu_X \sum_{k\in\mathbb{Z}} h_k$ and of ACF $r_{YY} = h * \widetilde{h} * r_{XX}$, where $\widetilde{h}_t = \overline{h_{-t}}$.*

*If in addition $r_{XX} \in l^1(\mathbb{Z})$, then the power spectral density of $Y_t$ is*

$$S_{YY}(\nu) = |H(\nu)|^2 S_{XX}(\nu). \tag{2.1}$$

## 1.2 Estimation of the mean and of the autocovariance function

Given the observation of a few samples of a WSS process, one may want to estimate its mean and its autocovariance function.

### 1.2.1 Parametric estimation

Let us first remind some basics of parametric estimation:

**Definition 4** (Estimator)**.** *Let $X$ be a complex random variable whose distribution is parameterized by $\theta$. An estimator $\widehat{\theta}$ of $\theta$ is a function of $X$. Its bias is defined as $b(\theta, \widehat{\theta}) = \mathbb{E}_\theta[\widehat{\theta}(X) - \theta]$, and its Mean Square Error (MSE) is defined as $R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\left[\left|\widehat{\theta}(X) - \theta\right|^2\right] = \mathrm{var}[\widehat{\theta}(X)] + |b(\theta, \widehat{\theta})|^2$.*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

There is a lower bound for $R(\theta, \widehat{\theta})$ called the *Cramér-Rao Lower Bound* (CRLB) for unbiased estimators [12] (the CRLB will be described in Chapter 5, Section 5.1).

**Definition 5** (Asymptotic approach of estimation)**.** *Let $X = [X_1, \ldots, X_N]^\top$ be a complex random vector whose distribution is parameterized by $\theta$. We say that an estimator $\widehat{\theta}_N$ of $\theta$ is* asymptotically unbiased *if* $\lim_{N\to+\infty} b(\theta, \widehat{\theta}_N) = 0$*. We say that it is* mean square consistent *if* $\lim_{N\mapsto+\infty} R(\theta, \widehat{\theta}_N) = 0$*.*

### 1.2.2 Estimation of the mean

**Definition 6** (Empirical mean)**.** *Let $X_t$ be a WSS process of mean $\mu_X$ and ACF $r_{XX}$. Given $N$ observed samples $X_1 \ldots X_N$, the empirical mean is defined as*

$$\hat{\mu}_X = \frac{1}{N}\sum_{t=1}^{N} X_t.$$

**Proposition 5** (Properties of the empirical mean)**.** *Let $X_t$ be a WSS process of mean $\mu_X$ and ACF $r_{XX}$. Then*

- *the empirical mean is an unbiased estimator: $\mathbb{E}[\hat{\mu}_X] = \mu_X$;*
- *its variance is* $\mathrm{var}[\hat{\mu}_X] = \frac{1}{N}\sum_{k=-(N-1)}^{N-1}\left(1 - \frac{|k|}{N}\right) r_{XX}(k)$*;*
- *If in addition $r_{XX} \in l^1(\mathbb{Z})$, then it is mean square consistent:* $\mathrm{var}[\hat{\mu}_X] \underset{N\to+\infty}{\sim} \frac{1}{N} S_{XX}(0)$*.*

### 1.2.3 Estimation of the autocovariance function

From now on, we consider that the mean $\mu_X$ of the WSS process $X_t$ is known, so without loss of generality we will assume that it is zero.

**Definition 7** (Empirical autocovariance function)**.** *Let $X_t$ be a centered WSS process of ACF $r_{XX}$. Given $N$ observed samples $X_1 \ldots X_N$, the empirical ACF is defined as $\forall 0 \le k < N$,*

$$\hat{r}_{XX}(k) = \frac{1}{N}\sum_{t=1}^{N-k} X_{t+k}\,\overline{X_t}, \tag{2.2}$$

*$\forall |k| \ge N$, $\hat{r}_{XX}(k) = 0$, and $\forall k \in \mathbb{Z}$, $\hat{r}_{XX}(-k) = \overline{\hat{r}_{XX}(k)}$.*

**Proposition 6** (Properties of the empirical autocovariance function)**.** *Let $X_t$ be a centered WSS process of ACF $r_{XX}$. Then:*

- *the empirical ACF $\hat{r}_{XX}$ is positive semi-definite;*
- *$\hat{r}_{XX}$ is an asymptotically unbiased estimator: $\mathbb{E}\left[\hat{r}_{XX}(k)\right] = \left(1 - \frac{|k|}{N}\right) r_{XX}(k)$;*
- *if in addition $X_t$ is a strong linear process (which means that $X_t = \sum_{k\in\mathbb{Z}} h_k Z_{t-k}$ where $h_k \in l^1(\mathbb{Z})$ and $Z_t \sim \mathrm{IID}(0, \sigma^2)$ with $\mathbb{E}\left[|Z_t|^4\right] < +\infty$), then $\hat{r}_{XX}$ is mean square consistent: $\forall k \in \mathbb{Z}$,* $\mathrm{var}\left[\hat{r}_{XX}(k)\right] = O\left(\frac{1}{N}\right)$*.*

# 2 Estimation of the power spectral density

We can now address the estimation of the power spectral density of a WSS process.

↪ For more details on the content of this section, we refer the reader to [2, chap. 2].

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Contexte académique } sans modifications
*Voir Page 98*

## 2.1 Periodogram

**Definition 8** (Periodogram and Correlogram)**.** *Let $N \in \mathbb{N}^*$. Let $X_t$ be a centered WSS process such that $r_{XX} \in l^1(\mathbb{Z})$. Its periodogram is defined as:*

$$\hat{S}_{P,XX}(\nu) = \frac{1}{N}\left|\sum_{t=1}^{N} X_t e^{-2\imath\pi\nu t}\right|^2. \tag{2.3}$$

*Its correlogram is defined as:*

$$\hat{S}_{C,XX}(\nu) = \sum_{k=-(N-1)}^{N-1} \hat{r}_{XX}(k) e^{-2\imath\pi\nu k}, \tag{2.4}$$

*where $\forall |k| \geq N$, $\hat{r}_{XX}(k) = 0$, $\forall k \in \mathbb{Z}$, $\hat{r}_{XX}(-k) = \overline{\hat{r}_{XX}(k)}$, and $\forall k \in [0, N[$, $\hat{r}_{XX}(k)$ is either:*

- *the unbiased estimator of the ACF:*

$$\hat{r}_{XX}(k) = \frac{1}{N-k}\sum_{t=1}^{N-k} X_{t+k}\overline{X_t};$$

- *or the biased (positive semi-definite) estimator of the ACF introduced in equation* (2.2)*:*

$$\hat{r}_{XX}(k) = \frac{1}{N}\sum_{t=1}^{N-k} X_{t+k}\overline{X_t}.$$

**Proposition 7** (Relationship between the periodogram and the correlogram)**.** *If $\hat{r}_{XX}$ is the biased estimator of the ACF, then the periodogram and the correlogram are equal: $\hat{S}_{P,XX}(\nu) = \hat{S}_{C,XX}(\nu)$.*

*Proof of Proposition 7.* First, we have

$$\begin{array}{rcl} \hat{S}_{P,XX}(\nu) & = & \frac{1}{N}\left|\sum_{t=1}^{N} X_t e^{-2\imath\pi\nu t}\right|^2 \\ & = & \frac{1}{N}\left(\sum_{t_1=1}^{N} X_{t_1} e^{-2\imath\pi\nu t_1}\right)\left(\sum_{t_2=1}^{N} \overline{X_{t_2}} e^{+2\imath\pi\nu t_2}\right) \\ & = & \sum_{t_1=1}^{N}\sum_{t_2=1}^{N} \frac{1}{N} X_{t_1}\overline{X_{t_2}} e^{-2\imath\pi\nu(t_1-t_2)}. \end{array}$$

Let $k = t_1 - t_2$. With $\hat{r}_{XX}(k)$ defined as in (2.2), we then get

$$\begin{array}{rcl} \hat{S}_{P,XX}(\nu) & = & \sum_{k=-(N-1)}^{N-1} \hat{r}_{XX}(k) e^{-2\imath\pi\nu k} \\ & = & \hat{S}_{C,XX}(\nu). \end{array}$$

□

## 2.2 Bias analysis of the periodogram

**Proposition 8.** *The mean of the periodogram defined in equation* (2.3) *is*

$$\mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = \sum_{k=-(N-1)}^{N-1}\left(1 - \frac{|k|}{N}\right) r_{XX}(k)\, e^{-2\imath\pi\nu k}. \tag{2.5}$$

*Proof of Proposition 8.* Let us consider the correlogram formulation of the periodogram in equation (2.4); we get

$$\mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = \sum_{k=-(N-1)}^{N-1} \mathbb{E}\left[\hat{r}_{XX}(k)\right] e^{-2\imath\pi\nu k}$$

where $\hat{r}_{XX}(k)$ is the biased estimator of the ACF, so that $\mathbb{E}\left[\hat{r}_{XX}(k)\right] = \left(1 - \frac{|k|}{N}\right) r_{XX}(k)$ as shown in Proposition 6. We thus get equation (2.5).

□

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

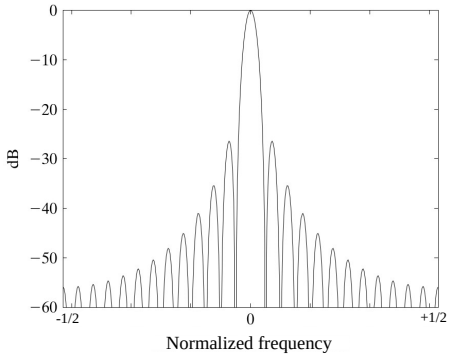



Figure 2.1: Fejer kernel, $W_B(\nu)/W_B(0)$, for $N = 25$.

Let us define the window function

$$w_B(k) = 1 - \frac{|k|}{N} \text{ if } |k| < N \text{ and } w_B(k) = 0 \text{ if } |k| \geq N, \tag{2.6}$$

whose expression appears in equation (2.5). Its Fourier transform is known as the *Fejer kernel*:

**Proposition 9** (Fejer kernel)**.** *The Fourier transform $W_B(\xi)$ of the window function $w_B(k)$ defined in* (2.6) *is called the* Fejer kernel*. Its closed-form expression is*

$$W_B(\xi) = \frac{1}{N}\left|\frac{\sin(\pi\xi N)}{\sin(\pi\xi)}\right|^2.$$

*Proof of Proposition 9.* Let $w_R(k) = \mathbf{1}_{[0,N-1]}(k)$. Then $w_B(k) = \frac{1}{N}(w_R * \tilde{w}_R)(k)$, therefore $W_B(\xi) = \frac{1}{N}|W_R(\xi)|^2$. However, $W_R(\xi) = \sum_{t=0}^{N-1} e^{-2\imath\pi\xi t} = \frac{1-e^{-2\imath\pi\xi N}}{1-e^{-2\imath\pi\xi}}$. Thus $|W_R(\xi)| = \left|\frac{\sin(\pi\xi N)}{\sin(\pi\xi)}\right|$. Therefore $W_B(\xi) = \frac{1}{N}\left|\frac{\sin(\pi\xi N)}{\sin(\pi\xi)}\right|^2$. □

Then equation (2.5) can be rewritten in the following form:

$$\mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = \sum_{k\in\mathbb{Z}} w_B(k) r_{XX}(k) e^{-2\imath\pi\nu k} = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{XX}(\nu-\xi) W_B(\xi) d\xi. \tag{2.7}$$

Equation (2.7) shows that $\mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right]$ can been seen as a local weighted average of the PSD, due to the convolution with the Fejer kernel. The Fejer kernel is represented in Figure 2.1. Considering equation (2.7), it permits us to interpret the bias of the periodogram as follows:

- the main lobe of the kernel induces smearing (its width is $\frac{2}{N}$);
- the side lobes of the kernel induce leakage;
- there is a loss of spectral resolution[1].

However, when $N \to +\infty$, the following proposition shows that the periodogram is asymptotically unbiased:

**Proposition 10.** *If $r_{XX} \in l^1(\mathbb{Z})$,, then $\hat{S}_{P,XX}(\nu)$ is asymptotically unbiased:* $\lim_{N\to+\infty} \mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = S_{XX}(\nu)$.

*Proof of Proposition 10.* Let $\forall N \in \mathbb{N}^*$, $\forall k \in \mathbb{Z}$, $u_{N,k} = \left(1 - \frac{|k|}{N}\right) r_{XX}(k)\, e^{-2\imath\pi\nu k}\, \mathbf{1}_{-(N-1),N-1}(k)$. Then equation (2.5) yields $\mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = \sum_{k\in\mathbb{Z}} u_{N,k}$. Let us now apply the dominated convergence theorem to the sequence $u_{N,k}$. First, let us check that the assumptions of the theorem hold:

[1] We say that two complex sinusoids are well-resolved when their frequencies are sufficiently distant from each other so as to produce two distinguishable spectral peaks in the periodogram. The spectral resolution is directly related to the width of the main lobe of the kernel. Do not confuse with the spectral *precision*, which is the distance between two successive frequency bins in the *Discrete Fourier Transform* (DFT).

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

- $\forall k \in \mathbb{Z}$, $u_{N,k} \underset{N\to+\infty}{\longrightarrow} r_{XX}(k)\, e^{-2\imath\pi\nu k}$;
- $\forall N \in \mathbb{N}^*$, $\forall k \in \mathbb{Z}$, $|u_{N,k}| \leq |r_{XX}(k)| \in l^1(\mathbb{Z})$.

Then the theorem proves that:

$$\lim_{N\to+\infty} \mathbb{E}\left[\hat{S}_{P,XX}(\nu)\right] = \lim_{N\to+\infty} \sum_{k\in\mathbb{Z}} u_{N,k} = \sum_{k\in\mathbb{Z}} \lim_{N\to+\infty} u_{N,k} = \sum_{k\in\mathbb{Z}} r_{XX}(k)\, e^{-2\imath\pi\nu k} = S_{XX}(\nu).$$

□

## 2.3 Variance analysis of the periodogram

Analyzing the variance of the periodogram is a more difficult problem than analyzing its bias. For this reason, we will first focus on a simple class of processes: the complex white noise.

**Definition 9** (Complex white noise)**.** *A complex (or circular) white noise is a complex WSS process such that* $\forall t, s \in \mathbb{Z}$,

$$\left\{ \begin{array}{r} \mathbb{E}[Z_t \overline{Z_s}] = \sigma^2 \delta_{t,s} \\ \mathbb{E}[Z_t Z_s] = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{r} \mathbb{E}[\mathcal{R}e(Z_t)\mathcal{R}e(Z_s)] = \frac{1}{2}\sigma^2\delta_{t,s} \\ \mathbb{E}[\mathcal{I}m(Z_t)\mathcal{I}m(Z_s)] = \frac{1}{2}\sigma^2\delta_{t,s} \\ \mathbb{E}[\mathcal{R}e(Z_t)\mathcal{I}m(Z_s)] = 0 \end{array} \right. .$$

In the case of the complex white noise, the asymptotic covariance between two values of the periodogram can easily be derived:

**Proposition 11** (Variance analysis of the periodogram of a complex white noise)**.** *If* $Z_t$ *is a Gaussian complex white noise of variance* $\sigma^2$*, then* $\forall \nu, \xi \in ]-\frac{1}{2}, \frac{1}{2}]$,

$$\lim_{N\to+\infty} \operatorname{cov}\left[\hat{S}_{P,ZZ}(\nu), \hat{S}_{P,ZZ}(\xi)\right] = \left\{ \begin{array}{cc} S_{ZZ}(\nu)^2 & \text{if } \nu = \xi \\ 0 & \text{if } \nu \neq \xi \end{array} \right. , \tag{2.8}$$

*where* $S_{ZZ}(\nu) = \sigma^2$ $\forall \nu \in \mathbb{R}$.

In order to prove this proposition, we will use the following lemma:

**Lemma 12** (Fourth order moments of Gaussian random variables)**.** *If* $a, b, c, d$ *are jointly Gaussian random variables, then*

$$\mathbb{E}[abcd] = \mathbb{E}[ab]\mathbb{E}[cd] + \mathbb{E}[ac]\mathbb{E}[bd] + \mathbb{E}[ad]\mathbb{E}[bc] - 2\mathbb{E}[a]\mathbb{E}[b]\mathbb{E}[c]\mathbb{E}[d]. \tag{2.9}$$

Lemma 12 is proved *e.g.* in [13]. Let us now prove Proposition 11:

*Proof of Proposition 11.* We have $\forall \nu, \xi \in ]-\frac{1}{2}, \frac{1}{2}]$,

$$\begin{aligned}
& \mathbb{E}\left[\hat{S}_{P,ZZ}(\nu)\hat{S}_{P,ZZ}(\xi)\right] \\
= \; & \mathbb{E}\left[\tfrac{1}{N}\left|\textstyle\sum_{t=1}^N Z_t e^{-2\imath\pi\nu t}\right|^2 \tfrac{1}{N}\left|\textstyle\sum_{t=1}^N Z_t e^{-2\imath\pi\xi t}\right|^2\right] \\
= \; & \tfrac{1}{N^2}\mathbb{E}\left[\textstyle\sum_{t_1,t_2,t_3,t_4=1}^N Z_{t_1} e^{-2\imath\pi\nu t_1}\overline{Z_{t_2}} e^{+2\imath\pi\nu t_2} Z_{t_3} e^{-2\imath\pi\xi t_3}\overline{Z_{t_4}} e^{+2\imath\pi\xi t_4}\right] \\
= \; & \tfrac{1}{N^2}\textstyle\sum_{t_1,t_2,t_3,t_4=1}^N e^{-2\imath\pi\nu(t_1-t_2)} e^{-2\imath\pi\xi(t_3-t_4)}\mathbb{E}\left[Z_{t_1}\overline{Z_{t_2}}Z_{t_3}\overline{Z_{t_4}}\right] \\
= \; & \tfrac{1}{N^2}\textstyle\sum_{t_1,t_2,t_3,t_4=1}^N e^{-2\imath\pi\nu(t_1-t_2)} e^{-2\imath\pi\xi(t_3-t_4)}\left(\mathbb{E}[Z_{t_1}\overline{Z_{t_2}}]\mathbb{E}[Z_{t_3}\overline{Z_{t_4}}] + \mathbb{E}[Z_{t_1}Z_{t_3}]\mathbb{E}[\overline{Z_{t_2}Z_{t_4}}] + \mathbb{E}[Z_{t_1}\overline{Z_{t_4}}]\mathbb{E}[\overline{Z_{t_2}}Z_{t_3}] - 2\mathbb{E}[Z_{t_1}]\mathbb{E}[\overline{Z_{t_2}}]\mathbb{E}[Z_{t_3}]\mathbb{E}[\overline{Z_{t_4}}]\right) \\
= \; & \tfrac{1}{N^2}\textstyle\sum_{t_1,t_2,t_3,t_4=1}^N e^{-2\imath\pi\nu(t_1-t_2)} e^{-2\imath\pi\xi(t_3-t_4)}\left(\sigma^2\delta_{t_1,t_2}\sigma^2\delta_{t_3,t_4} + 0 + \sigma^2\delta_{t_1,t_4}\sigma^2\delta_{t_2,t_3} - 0\right) \\
= \; & \tfrac{\sigma^4}{N^2}\textstyle\sum_{t_1,t_2,t_3,t_4=1}^N e^{-2\imath\pi\nu(t_1-t_2)} e^{-2\imath\pi\xi(t_3-t_4)}\left(\delta_{t_1,t_2}\delta_{t_3,t_4} + \delta_{t_1,t_4}\delta_{t_2,t_3}\right) \\
= \; & \tfrac{\sigma^4}{N^2}\left(N^2 + \textstyle\sum_{t_1,t_2=1}^N e^{-2\imath\pi(\nu-\xi)(t_1-t_2)}\right) \\
= \; & \tfrac{\sigma^4}{N^2}\left(N^2 + \left(\textstyle\sum_{t_1=1}^N e^{-2\imath\pi(\nu-\xi)t_1}\right)\left(\textstyle\sum_{t_2=1}^N e^{+2\imath\pi(\nu-\xi)t_2}\right)\right) \\
= \; & \tfrac{\sigma^4}{N^2}\left(N^2 + \left|\tfrac{\sin(\pi(\nu-\xi)N)}{\sin(\pi(\nu-\xi))}\right|^2\right) \\
= \; & \sigma^4 + \sigma^4\left|\tfrac{\sin(\pi(\nu-\xi)N)}{N\sin(\pi(\nu-\xi))}\right|^2 .
\end{aligned}$$

**Roland Badeau and François Roueff**
**roland.badeau@telecom-paris.fr**

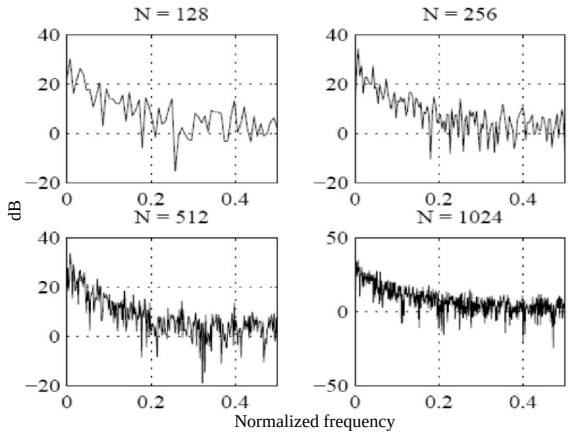



Figure 2.2: Periodogram of a WSS process for $N = 128, 256, 512, 1024$.

In particular,

$$\lim_{N\to+\infty} \mathbb{E}\left[\hat{S}_{P,ZZ}(\nu)\hat{S}_{P,ZZ}(\xi)\right] = \begin{cases} 2\sigma^4 & \text{if } \nu = \xi \\ \sigma^4 & \text{if } \nu \neq \xi \end{cases}$$

In addition, Proposition 10 shows that

$$\lim_{N\to+\infty} \mathbb{E}\left[\hat{S}_{P,ZZ}(\nu)\right] = S_{ZZ}(\nu) = \sigma^2.$$

Since $\operatorname{cov}\left[\hat{S}_{P,ZZ}(\nu), \hat{S}_{P,ZZ}(\xi)\right] = \mathbb{E}\left[\hat{S}_{P,ZZ}(\nu)\hat{S}_{P,ZZ}(\xi)\right] - \mathbb{E}\left[\hat{S}_{P,ZZ}(\nu)\right]\mathbb{E}\left[\hat{S}_{P,ZZ}(\xi)\right]$, we finally get

$$\lim_{N\to+\infty} \operatorname{cov}\left[\hat{S}_{P,ZZ}(\nu), \hat{S}_{P,ZZ}(\xi)\right] = \begin{cases} \sigma^4 & \text{if } \nu = \xi \\ 0 & \text{if } \nu \neq \xi \end{cases}$$

□

Actually, Proposition 11 also holds in the more general case of a WSS process obtained by filtering a Gaussian white noise, as shown in the following Proposition.

**Proposition 13** (Variance analysis of the periodogram of a stationary Gaussian process)**.** *Let $Z_t$ be a complex Gaussian white noise, and let*

$$X_t = \sum_{k=1}^{\infty} h_k Z_{t-k} \tag{2.10}$$

*where $(k\, h_k)_k \in l^1(\mathbb{N}^*)$. Then $\forall \nu, \xi \in ]-\frac{1}{2}, \frac{1}{2}]$,*

$$\lim_{N\to+\infty} \operatorname{cov}\left[\hat{S}_{P,XX}(\nu), \hat{S}_{P,XX}(\xi)\right] = \begin{cases} S_{XX}(\nu)^2 & \text{if } \nu = \xi \\ 0 & \text{if } \nu \neq \xi \end{cases} \tag{2.11}$$

Proposition 13 shows that $\hat{S}_{P,XX}$ is not even asymptotically mean square consistent, since $\lim_{N\to+\infty} \operatorname{var}\left[\hat{S}_{P,XX}(\nu)\right] = S_{XX}(\nu)^2 \neq 0$ (as illustrated in Figure 2.2).

The following proof of Proposition 13 is technical and is provided only for the interested reader:

*Proof of Proposition 13.* The PSD of the WSS process $X_t$ defined in (2.10) is given by

$$S_{XX}(\nu) = |H(\nu)|^2 S_{ZZ}(\nu) \tag{2.12}$$

(cf. equation (2.1)). Here, $H(\nu) = \sum_{k=1}^{\infty} h_k e^{-2\imath\pi\nu k}$. We will omit the index $P$ of $\hat{S}_{P,XX}(\nu)$ in order to simplify the notation. We will show that the periodograms approximately satisfy an equation of the form of (2.12) that is satisfied by the true PSDs:

$$\text{For } N \gg 1,\ \hat{S}_{XX}(\nu) = |H(\nu)|^2 \hat{S}_{ZZ}(\nu) + O\left(\tfrac{1}{\sqrt{N}}\right). \tag{2.13}$$

Let us prove (2.13): with the change of variable $p = t - k$, we get

$$\begin{aligned}
\frac{1}{\sqrt{N}} \sum_{t=1}^{N} X_t e^{-2\imath\pi\nu t} &= \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \sum_{k=1}^{\infty} h_k Z_{t-k} e^{-2\imath\pi\nu(t-k)} e^{-2\imath\pi\nu k} \\
&= \frac{1}{\sqrt{N}} \sum_{k=1}^{\infty} h_k e^{-2\imath\pi\nu k} \sum_{p=1-k}^{N-k} Z_p e^{-2\imath\pi\nu p} \\
&= \frac{1}{\sqrt{N}} \sum_{k=1}^{\infty} h_k e^{-2\imath\pi\nu k} \\
&\quad \cdot \left[ \sum_{p=1}^{N} Z_p e^{-2\imath\pi\nu p} + \sum_{p=1-k}^{0} Z_p e^{-2\imath\pi\nu p} - \sum_{p=N-k+1}^{N} Z_p e^{-2\imath\pi\nu p} \right] \\
&= H(\nu) \left[ \frac{1}{\sqrt{N}} \sum_{p=1}^{N} Z_p e^{-2\imath\pi\nu p} \right] + \rho(\nu)
\end{aligned} \tag{2.14}$$

where

$$\rho(\nu) = \frac{1}{\sqrt{N}} \sum_{k=1}^{\infty} h_k e^{-2\imath\pi\nu k} \varepsilon_k(\nu)$$

with $\varepsilon_k(\nu) = \sum_{p=1-k}^{0} Z_p e^{-2\imath\pi\nu p} - \sum_{p=N-k+1}^{N} Z_p e^{-2\imath\pi\nu p}$.

Next, note that

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_k(\nu)\right] &= 0, \\
\mathbb{E}\left[\varepsilon_k(\nu)\varepsilon_j(\nu)\right] &= 0 \text{ for all } k \text{ and } j, \text{ and} \\
\mathbb{E}\left[\varepsilon_k(\nu)\overline{\varepsilon_j(\nu)}\right] &= 2\sigma^2 \min(k, j)
\end{aligned}$$

which imply

$$\mathbb{E}[\rho(\nu)] = 0, \quad \mathbb{E}\left[\rho^2(\nu)\right] = 0$$

and

$$\begin{aligned}
\mathbb{E}\left[|\rho(\nu)|^2\right] &= \frac{1}{N} \left| \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} h_k e^{-2\imath\pi\nu k} \overline{h_j} e^{2\imath\pi\nu j} \mathbb{E}\left[\varepsilon_k(\nu)\overline{\varepsilon_j(\nu)}\right] \right| \\
&= \frac{2\sigma^2}{N} \left| \sum_{k=1}^{\infty} h_k e^{-2\imath\pi\nu k} \left\{ \sum_{j=1}^{k} \overline{h_j} e^{2\imath\pi\nu j} j + \sum_{j=k+1}^{\infty} \overline{h_j} e^{2\imath\pi\nu j} k \right\} \right| \\
&\le \frac{2\sigma^2}{N} \sum_{k=1}^{\infty} |h_k| \left\{ \sum_{j=1}^{\infty} |h_j| j + \sum_{j=1}^{\infty} |h_j| k \right\} \\
&= \frac{4\sigma^2}{N} \left( \sum_{k=1}^{\infty} |h_k| \right) \left( \sum_{j=1}^{\infty} |h_j| j \right)
\end{aligned}$$

Since $\sum_{k=1}^{\infty} k\, h_k$ is finite, we have

$$\mathbb{E}\left[|\rho(\nu)|^2\right] \le \frac{\text{constant}}{N}. \tag{2.15}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Now, from the squared modulus of (2.14), we obtain

$$\hat{S}_{XX}(\nu) = |H(\nu)|^2 \hat{S}_{ZZ}(\nu) + \gamma(\nu)$$

where

$$\gamma(\nu) = \overline{H(\nu)E(\nu)}\rho(\nu) + H(\nu)E(\nu)\overline{\rho(\nu)} + \rho(\nu)\overline{\rho(\nu)}$$

and where

$$E(\nu) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} Z_t e^{-2\imath\pi\nu t}.$$

Both $E(\nu)$ and $\rho(\nu)$ are linear combinations of Gaussian random variables, so they are also Gaussian distributed. This means that the fourth-order moment formula (2.9) can be used to obtain the second-order moment of $\gamma(\nu)$. By doing so, and also by using (2.15) and the fact that, for example,

$$\begin{aligned}\left|\mathbb{E}\left[\rho(\nu)\overline{E(\nu)}\right]\right| &\leq \left[\mathbb{E}\left[|\rho(\nu)|^2\right]\right]^{1/2} \left[\mathbb{E}\left[|E(\nu)|^2\right]\right]^{1/2} \\ &= \frac{\text{constant}}{\sqrt{N}} \cdot \left[\mathbb{E}\left[\left|\hat{S}_{ZZ}(\nu)\right|^2\right]\right]^{1/2} = \frac{\text{constant}}{\sqrt{N}},\end{aligned}$$

we can verify that $\gamma(\nu) = O(1/\sqrt{N})$, and hence the proof of (2.13) is concluded. The main result of Proposition 13 is derived by combining (2.8) and (2.13), which proves (2.11).

□

## 2.4 Blackman-Tukey method

In order to reduce the variance of the periodogram, the Blackman-Tukey method consists in truncating the empirical ACF. Let $X_t$ be a centered WSS process such that $r_{XX} \in l^1(\mathbb{Z})$. With $M < N$, the Blackman-Tukey estimator of its PSD is defined as:

$$\hat{S}_{BT,XX}(\nu) = \sum_{k=-M+1}^{M-1} \hat{r}_{XX}(k)\, e^{-2\imath\pi\nu k}.$$

The Blackman-Tukey estimator satisfies the following properties (*cf.* [2, chap. 2, sec. 2.5]):

- if $M \to +\infty$, $\hat{S}_{BT,XX}$ is asymptotically unbiased;
- if $M/N \to 0$, $\text{var}\left(\hat{S}_{BT,XX}(\nu)\right) = O\left(\frac{M}{N}\right) \to 0$;
- in particular, if $M = N^\alpha$ with $0 < \alpha < 1$, $\hat{S}_{BT,XX}$ is mean square consistent.

The Blackman-Tukey method offers a trade-off between the spectral resolution (which is $O(\frac{1}{M})$) and the variance (which is $O\left(\frac{M}{N}\right)$).

This approach can be further improved by introducing a symmetric window $w$ of support $[-(M-1),(M-1)]$ with $M < N$, such that $w(0) = 1$. The Blackman-Tukey estimator can then be seen as a windowed periodogram:

$$\hat{S}_{BT,XX}(\nu) = \sum_{k=-(M-1)}^{M-1} w(k)\widehat{r}_{XX}(k)\, e^{-2\imath\pi\nu k}.$$

Then

$$\hat{S}_{BT,XX}(\nu) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{S}_{P,XX}(\nu-\xi)W(\xi)d\xi \tag{2.16}$$

where $W(\xi) = \sum_{k=-(M-1)}^{M-1} w_k e^{-2\imath\pi\xi k} \in \mathbb{R}$: the Blackman-Tukey method performs a local weighted average of the spectrum. Note that if in addition $w(k)$ is positive semidefinite, then $W(\xi) \geq 0$, therefore $\hat{S}_{BT,XX}(\nu) \geq 0\ \forall \nu \in \mathbb{R}$.

The choice of the window's length is based on a trade-off between spectral resolution and variance. The selection of the window's shape is based on a trade-off between smearing and leakage effects (see Figure 2.3):

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

- the rectangular window (see Figure 2.3a) has width $\frac{1}{M}$, second lobe at -13 *Decibels* (dB), and decrease of -6 dB/octave;
- the Bartlett window (see Figure 2.3b) has width $\frac{2}{M}$, second lobe at -26 dB, and decrease of -12 dB/octave;
- the Hann window (see Figure 2.3c) has width $\frac{2}{M}$, second lobe at -31 dB, and decrease of -18 dB/octave;
- the Hamming window (see Figure 2.3d) has width $\frac{2}{M}$, second lobe at -41 dB, and decrease of -6 dB/octave;
- the Blackman window (see Figure 2.3e) has width $\frac{3}{M}$, second lobe at -57 dB, and decrease of -18 dB/octave.

## 2.5 Bartlett method

Compared to the original periodogram, the Bartlett method splits up the data and averages several periodograms in order to reduce the variance. The $N$ signal samples are segmented into $L$ sub-samples of size $M = \frac{N}{L}$. The Bartlett estimator is then defined as

$$\hat{S}_{B,XX}(\nu) = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{M}\left|\sum_{t=1}^{M}\tilde{X}_{i,t}e^{-2\imath\pi\nu t}\right|^2,$$

where $\tilde{X}_{i,t} = X_{(i-1)M+t}$ for $t \in [1, M]$ and $i \in [1, L]$.

Its spectral resolution is shown to be $O(\frac{1}{M})$ and its variance is $O(\frac{M}{N})$ (*cf.* [2, chap. 2, sec. 2.7.1]). The Bartlett method offers the same trade-off between spectral resolution and variance as the Blackman-Tukey estimate with a rectangular window.

## 2.6 Welch method

The Welch method is a refinement of the Bartlett method, which introduces two novelties: the data segments overlap, and each data segment is windowed: $\tilde{X}_{i,t} = X_{(i-1)K+t}$ for $t \in [1, M]$ and $i \in [1, S]$.

If $K = M$, the Welch method is equivalent to the Bartlett method: $S = L = \frac{N}{M}$. However the recommended choice is $K = \frac{M}{2}$, $S \approx \frac{2N}{M}$.

The Welch estimator is thus defined as $\hat{S}_{W,XX}(\nu) = \frac{1}{S}\sum_{i=1}^{S}\hat{S}^{(i)}_{P,XX}(\nu)$ with $\hat{S}^{(i)}_{P,XX}(\nu) = \frac{1}{MP}\left|\sum_{t=1}^{M} v(t)\tilde{X}_{i,t}e^{-2\imath\pi\nu t}\right|^2$, where $v(t)$ denotes the window, and $P = \frac{1}{N}\sum_{t=1}^{M}|v(t)|^2$ in order to normalize every periodogram. This method offers a better control of the smearing and leakage, but its variance is similar to that of the Bartlett method (*cf.* [2, chap. 2, sec. 2.7.2]).

## 2.7 Daniell method

The Daniell method aims to reduce the variance by smoothing the periodogram:

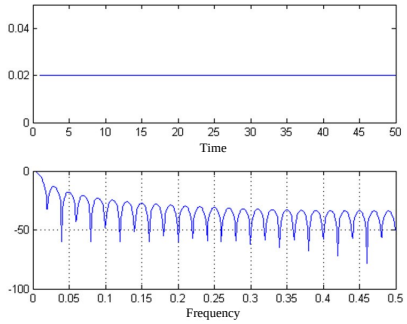
$$\hat{S}_{D,XX}(\nu) = \frac{1}{2J+1}\sum_{j=-J}^{J}\hat{S}_{P,XX}\left(\nu + \frac{j}{\tilde{N}}\right),$$

where $\tilde{N} = N$ without zero-padding, or $\tilde{N} > N$ with zero-padding. The continuous version of this method is

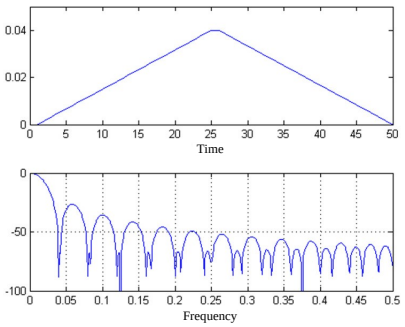
$$\hat{S}_{D,XX}(\nu) = \frac{1}{\beta}\int_{\nu-\frac{\beta}{2}}^{\nu+\frac{\beta}{2}}\hat{S}_{P,XX}(\nu+\xi)\,d\xi,$$

with $\beta = \frac{2J}{\tilde{N}}$. Thus the Daniell method can be seen as a particular case of the Blackman-Tukey method, with $W(\xi) = \frac{1}{\beta}$ if $\xi \in [-\frac{\beta}{2}, \frac{\beta}{2}]$, or $W(\xi) = 0$ otherwise (*cf.* equation (2.16), and [2, chap. 2, sec. 2.7.3]).

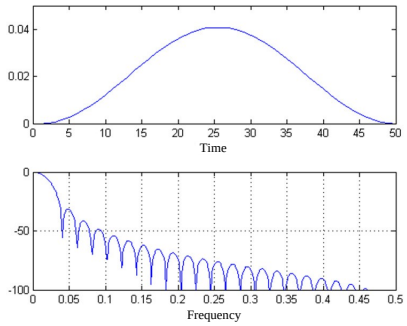
**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr



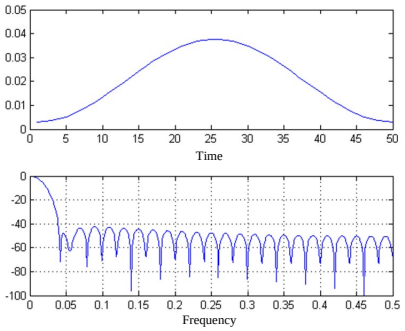

(a) Rectangular window:
$w(k) = \mathbf{1}_{[-(M-1)\ldots M-1]}(k)$.



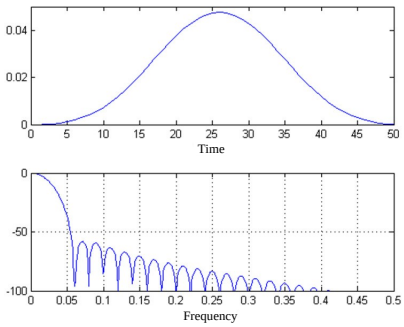

(b) Bartlett window:
$w(k) = \left(1 - \frac{|k|}{M}\right) \mathbf{1}_{[-(M-1)\ldots M-1]}(k)$.




(c) Hann window:
$w(k) = \left(0.5 + 0.5\cos(\frac{\pi k}{M})\right) \mathbf{1}_{[-(M-1)\ldots M-1]}(k)$.




(d) Hamming window:
$w(k) = \left(0.54 + 0.46\cos(\frac{\pi k}{M-1})\right) \mathbf{1}_{[-(M-1)\ldots M-1]}(k)$.




(e) Blackman window:
$w(k) = \left(0.4266 + 0.4965\cos(\frac{\pi k}{M-1}) + 0.076\cos(\frac{2\pi k}{M-1})\right) \mathbf{1}_{[-(M-1)\ldots M-1]}(k)$.

Figure 2.3: Common window shapes

# Chapter 3

# Parametric estimation of rational spectra

The principal difference between the spectral estimation methods of Chapter 2 and those in this chapter, is that in Chapter 2 we made no assumption on the studied signal (except for its stationarity). The parametric or model-based methods of spectral estimation assume that the signal satisfies a generating model with known functional form, and then proceed by estimating the parameters in the assumed model. The signal's spectral characteristics of interest are then derived from the estimated model. In those cases where the assumed model is a close approximation to the reality, it is no wonder that the parametric methods provide more accurate spectral estimates than the non-parametric techniques. The non-parametric approach to PSD estimation remains useful, though, in applications where there is little or no information about the signal in question. In this chapter, we discuss parametric methods for rational spectra which form a dense set in the class of continuous spectra; more precisely, we discuss methods for estimating the parameters in rational spectral models. Furthermore, parametric methods provide consistent spectral estimates and hence (for large sample sizes, at least), the issue of statistical behavior is not critical.

This chapter is structured as follows: in Section 1, we first remind the notion of linear processes, and two particular cases of interest: the *Moving Average* (MA) and the *Autoregressive* (AR) models. Then in Section 2, we introduce the maximum entropy spectral estimator, and we show that it is equivalent to AR modeling. Thus in Section 3, we remind the linear prediction approach for estimating AR models, the expression of Yule-Walker equations, we prove the stability of the causal AR filter obtained by solving these equations, and we remind the Levinson-Durbin algorithm that efficiently computes the solution. Then, in Section 4, we remind known properties of *Autoregressive Moving Average* (ARMA) models. Finally, in Section 5, we introduce the Durbin method for estimating ARMA models.

↪ For more details on the content of this chapter, we refer the reader to [2, chap. 3].

## 1 Reminder: linear processes

In this section, we first remind the notion of linear processes, and two particular cases of interest: the MA and the AR models and their specific properties.

### 1.1 Linear processes

**Definition 10** (Linear process)**.** $(X_t)_{t\in\mathbb{Z}}$ *is a* linear process *if and only if there is* $\mu_X \in \mathbb{C}$, $Z_t \sim \mathrm{WN}(0, \sigma^2)$ *and* $(h_n)_{n\in\mathbb{Z}} \in l_1(\mathbb{Z})$ *such that* $\forall t \in \mathbb{Z}$, $X_t = \mu_X + \sum_{n=-\infty}^{+\infty} h_n Z_{t-n}$. *In addition,*

- $(X_t)_{t\in\mathbb{Z}}$ *is* causal *with respect to* $(Z_t)_{t\in\mathbb{Z}}$ *if and only if* $h_n = 0\ \forall n < 0$*;*
- $(X_t)_{t\in\mathbb{Z}}$ *is* invertible *with respect to* $(Z_t)_{t\in\mathbb{Z}}$ *if and only if there is a sequence* $(g_n)_{n\geq 0} \in l_1(\mathbb{N})$ *such that* $Z_t = \sum_{n=0}^{+\infty} g_n(X_{t-n} - \mu_X)\ \forall t \in \mathbb{Z}$.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Proposition 14** (Properties of linear processes)**.** *Let $X_t$ be a linear process as defined in Definition 10. Then $X_t$ is a WSS process of mean $\mu_X$, ACF $r_{XX}(k) = \mathbb{E}[(X_{t+k} - \mu_X)(\overline{X_t} - \overline{\mu_X})] = \sigma^2 \sum_{n=-\infty}^{+\infty} h_{n+k}\overline{h}_n$, and spectral density $S_{XX}(\nu) = \sigma^2|H(\nu)|^2$ with $H(\nu) = \sum_{k\in\mathbb{Z}} h_k e^{-2\imath\pi\nu k}$.*

*Proof of Proposition 14.* This proposition is a consequence of the filtering theorem for WSS processes (Theorem 4 page 11). □

## 1.2 Moving average processes

MA processes are linear processes such that filter $h$ has a *Finite Impulse Response* (FIR).

### 1.2.1 MA($q$) process

**Definition 11** (MA($q$) process)**.** *The process $(X_t)_{t\in\mathbb{Z}}$ is* moving average *of order $q$ (or MA($q$)) if and only if $X_t = \sum_{n=0}^{q} b_n Z_{t-n}$ where $Z_t \sim \mathrm{WN}(0, \sigma^2)$, $b_n \in \mathbb{C}$ and $b_0 = 1$.*

**Proposition 15** (Properties of MA processes)**.** *Let $X_t$ be an MA process as defined in Definition 11. Then $X_t$ is a WSS process of mean 0, of ACF $r_{XX}(k) = \sigma^2 \sum_{n=0}^{q-k} b_{n+k}\overline{b}_n$ if $0 \le k \le q$ and $r_{XX}(k) = 0$ if $k > q$, and of spectral density $S_{XX}(\nu) = \sigma^2 \left|\sum_{n=0}^{q} b_n e^{-2\imath\pi\nu n}\right|^2$.*

*Proof of Proposition 15.* This proposition is a consequence of the filtering theorem for WSS processes (Theorem 4 page 11). □

The PSD of an MA(1) process is illustrated in Figure 3.1.

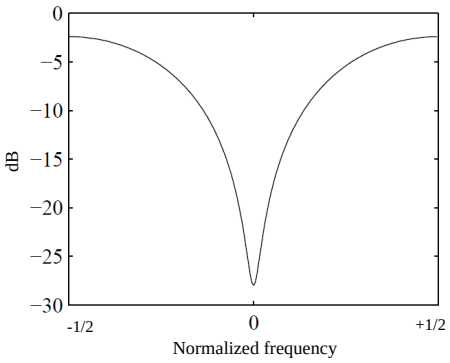



Figure 3.1: PSD (in dB) of an MA(1) process with $\sigma = 1$ and $b_1 = -0.9$

### 1.2.2 Characterization of an MA($q$) process

**Theorem 16** (Characterization of an MA($q$) process)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a centered WSS process of ACF $r_{XX}(k)$, and let $q \ge 1$. Then the two following assertions are equivalent:*

- *$X_t$ is an MA process of minimal order $q$;*
- *$r_{XX}(q) \neq 0$ and $r_{XX}(k) = 0\ \forall k \ge q+1$.*

**Corollary 17.** *The sum of two uncorrelated MA($q$) processes is an MA($q$) process.*

## 1.3 Autoregressive processes

AR processes are linear processes such that filter $h$ is AR.

**Definition 12** (AR($p$) process)**.** *The process $(X_t)_{t\in\mathbb{Z}}$ is* autoregressive *of order $p$ (or AR($p$)) if and only if it is WSS and solution of the equation*

$$X_t = Z_t + \sum_{n=1}^{p} a_n X_{t-n} \tag{3.1}$$

*where $Z_t \sim \mathrm{WN}(0, \sigma^2)$, and $a_n \in \mathbb{C}$.*

The existence and unicity of a WSS solution to equation (3.1) is a difficult question, which did not exist for MA processes.

**Example 1** (AR(1) process)**.** *Let $X_t$ be an AR($p$) process with $p = 1$. Then*

- *if $|a_1| < 1$, we have $X_t = \sum_{n=0}^{+\infty} a_1^n Z_{t-n}$ (convergence in $L^2$ and a.s.);*
- *if $|a_1| > 1$, we have $X_t = -\sum_{n=1}^{+\infty} a_1^{-n} Z_{t+n}$ (convergence in $L^2$ and a.s.);*
- *in both cases, the filtering theorem for WSS processes (Theorem 4 page 11) proves that if $|a_1| \neq 1$, $X_t$ is WSS of mean 0 and of spectral density $S_{XX}(\nu) = \frac{\sigma^2}{|1-a_1 e^{-2i\pi\nu}|^2}$.*
- *if $|a_1| = 1$, the recursive equation (3.1) does not admit any WSS solution.*

Three trajectories of a causal Gaussian AR(1) process are represented in Figure 3.2.
The PSD of a causal Gaussian AR(1) process is illustrated in Figure 3.3.

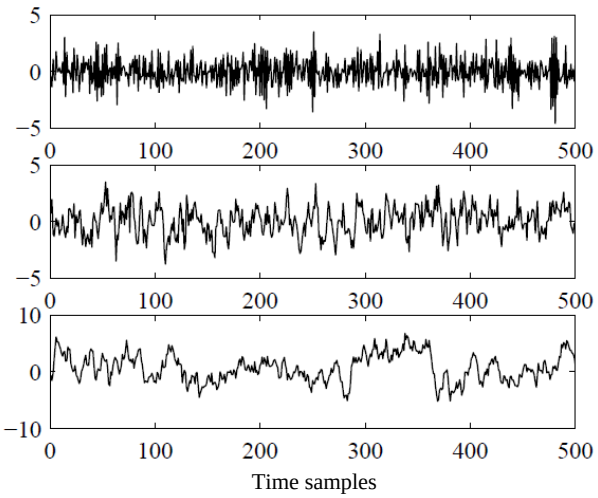



Figure 3.2: Trajectories of a Gaussian AR(1) process, of length 500. Top: $a_1 = -0.7$. Center: $a_1 = 0.5$. Bottom: $a_1 = 0.9$.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

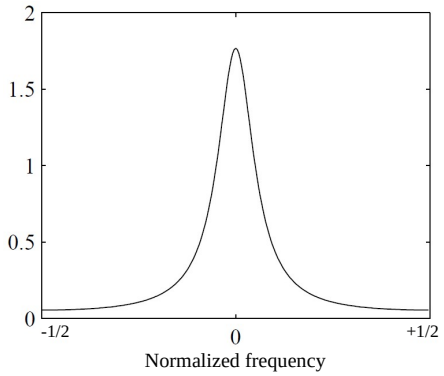



Figure 3.3: PSD of a Gaussian AR(1) process, with $\sigma = 1$ and $a_1 = 0.7$.

Let us now address the general case ($p \in \mathbb{N}^*$):

**Proposition 18** (Properties of AR processes)**.** *There is a WSS solution to equation* (3.1) *in Definition 12 if and only if* $A(z) \triangleq 1 - \sum_{n=1}^{p} a_n z^{-n} \neq 0$ *for* $|z| = 1$. *Then with* $\frac{1}{A(z)} = \sum_{n=-\infty}^{+\infty} h_n z^{-n}$ *where* $\sum_{n\in\mathbb{Z}} |h_n| < +\infty$, *we get* $X_t = \sum_{n\in\mathbb{Z}} h_n Z_{t-n}$. *The AR process* $X_t$ *has mean* 0, *ACF* $r_{XX}(k) = \sigma^2 \sum_{n=-\infty}^{+\infty} h_{n+k}\overline{h}_n$, *and PSD* $S_{XX}(\nu) = \frac{\sigma^2}{\left|1-\sum_{n=1}^{p} a_n e^{-2\imath\pi\nu n}\right|^2}$. *If all the zeros of* $A(z)$ *are strictly inside the unit circle, the solution is causal; if all the zeros of* $A(z)$ *are strictly outside the unit circle, the solution is anti-causal; otherwise,* $X_t$ *is a* mixed *AR process.*

*Proof of Proposition 18.* This proposition is a consequence of the filtering theorem for WSS processes (Theorem 4 page 11). □

# 2 Maximum entropy spectral estimation

In this section, we introduce the maximum entropy spectral estimator, and we show that it is equivalent to AR modeling.

↪ For more details on the content of this section, we refer the reader to [14, pp. 16–21].

Let $X_t$ be a centered WSS process such that $r_{XX} \in l^1(\mathbb{Z})$. In the context of non-parametric spectral estimation, the PSD $S_{XX}(\nu)$ is estimated from $N$ samples $X_1 \dots X_N$. For a given $M \leq N$, the periodogram method and the Blackman-Tukey methods, presented in Section 2 page 12, consist of:

- first computing estimates $\hat{r}_{XX}(k)$ of $r_{XX}(k)$ for $k \in [-M+1, M-1]$;
- then estimating $\hat{S}_{XX}(\nu)$ via a (windowed) DTFT of $\hat{r}_{XX}$.

Considering the characterization of MA processes in Theorem 16, we can remark that these methods actually approximate the true PSD by that of an MA process.

In comparison, the maximum entropy spectral estimator relies on a new idea: with fixed $\hat{r}_{XX}(k) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \hat{S}_{XX}(\nu) e^{+2\imath\pi\nu k} d\nu$ for $k \in [-M+1, M-1]$, it computes the estimate $\hat{S}_{XX}(\nu)$ that maximizes the *entropy* of the WSS probability distribution. This is a *blind* estimation approach: no information is available about the WSS process beyond the knowledge of $\hat{r}_{XX}(k)$ for $k \in [-M+1, M-1]$. In particular, this approach does not implicitly assume that $X_t$ is an MA process.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

## 2.1 Entropy of a Gaussian random vector

Let us now introduce the notion of entropy.

↪ For more details on the content of this section, we refer the reader to [14, chap. 2 p. 67].

For a discrete random variable with $M$ values, the entropy is defined as

$$H = \sum_{k=1}^{M} p_k \log_2\left(\tfrac{1}{p_k}\right) = \frac{1}{\ln(2)} \sum_{k=1}^{M} p_k \ln\left(\tfrac{1}{p_k}\right).$$

It can be easily proved that the entropy is always non-negative, and it measures the "uncertainty" in the outcome of the random variable. In the continuous case, for $N$ real-valued of complex-valued random variables $X_1, \ldots, X_N$, the entropy is defined as

$$H_N = -\int p(x_1, \ldots, x_N) \ln(p(x_1, \ldots, x_N) c^{\frac{N}{2}}) dx_1 \ldots dx_N \tag{3.2}$$

where $c$ is a normalizing constant. In this case, the non-negativity is no longer guaranteed, but the entropy still measures the uncertainty in the outcome of the random vector $[X_1, \ldots, X_N]^\top$. Now it is well known that if the random variables $X_1, \ldots, X_N$ are jointly Gaussian, then their probability density function (PDF) is well-defined when the $N$-dimensional covariance matrix $\boldsymbol{R}_N$ of the random vector $[X_1, \ldots, X_N]^\top$ is non-singular:

$$p(x_1, \ldots, x_N) = \frac{1}{(2\pi)^{\frac{N}{2}} \det(\boldsymbol{R}_N)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_X)^\top \boldsymbol{R}_N^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X)\right). \tag{3.3}$$

**Proposition 19.** *If $[X_1, \ldots, X_N]^\top$ is a Gaussian random vector of covariance matrix $\boldsymbol{R}_N$, then with $c = 2\pi e$, the entropy defined in equation* (3.2) *admits a simple closed form:*

$$H_N = \frac{1}{2} \ln(\det(\boldsymbol{R}_N)). \tag{3.4}$$

*Proof of Proposition 19.* Equation (3.2) yields

$$H_N = -\mathbb{E}[\ln(p(x_1, \ldots, x_N)] - \frac{N}{2} \ln(c)$$

where, due to equation (3.3),

$$\ln(p(x_1, \ldots, x_N)) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\boldsymbol{R}_N)) - \frac{1}{2} \mathrm{trace}\left(\boldsymbol{R}_N^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X)(\boldsymbol{x} - \boldsymbol{\mu}_X)^\top\right).$$

Therefore

$$H_N = \frac{N}{2} \ln(2\pi e) + \frac{1}{2} \ln(\det(\boldsymbol{R}_N)) - \frac{N}{2} \ln(c).$$

With $c = 2\pi e$, we thus get equation (3.4). □

## 2.2 Entropy of a Gaussian stationary process

Now, we are interested in defining the entropy of a Gaussian stationary process, i.e. a Gaussian random vector of infinite dimension. The problem is that $H_N$ as defined above diverges when $N \to +\infty$. We will thus consider instead the *entropy rate*, which is well-defined [14, chap 2, p. 16]:

**Definition 13** (Entropy rate)**.** *For a Gaussian stationary process, the* entropy rate *is*

$$H = \lim_{N \to +\infty} \frac{H_N}{N} = \lim_{N \to +\infty} \frac{1}{2} \ln((\det(\boldsymbol{R}_N))^{\frac{1}{N}}).$$

In order to study the limit in Definition 13, we will need the following theorem, that will be admitted:

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Theorem 20** (Szegö theorem)**.** *Let $X_t$ be a WSS process, $\boldsymbol{R}_N$ be the covariance matrix of the random vector $[X_1 \dots X_N]^\top$, $\sigma_0^2 \dots \sigma_{N-1}^2$ be the eigenvalues of $\boldsymbol{R}_N$, and $g$ be any continuous function. Then*

$$\lim_{N\to+\infty} \frac{1}{N}(g(\sigma_0^2) + \dots + g(\sigma_{N-1}^2)) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} g(S_{XX}(\nu))d\nu.$$

**Corollary 21.** *The entropy rate introduced in Definition 13 admits the following closed form expression:*

$$H = \frac{1}{2}\int_{-\frac{1}{2}}^{+\frac{1}{2}} \ln(S_{XX}(\nu))d\nu. \tag{3.5}$$

*Proof.* Wee apply Theorem 20 to function $g(.) = \ln(.)$. Since $\ln((\det(\boldsymbol{R}_N)) = \ln(\sigma_0^2) + \dots + \ln(\sigma_{N-1}^2)$, we have $H = \frac{1}{2}\lim_{N\to+\infty} \frac{1}{N}(g(\sigma_0^2) + \dots + g(\sigma_{N-1}^2))$. Thus the Szegö theorem proves that $H = \frac{1}{2}\int_{-\frac{1}{2}}^{+\frac{1}{2}} \ln(S_{XX}(\nu))d\nu$. □

## 2.3 Maximum entropy method

We can now state the main result of this section:

**Proposition 22.** *Among all Gaussian stationary processes with fixed ACF $\hat{r}_{XX}(k)$ for $|k| < M$, the one which maximizes the entropy rate is the AR($M-1$) process.*

*Proof.* Let $r(k)$ be the autocovariance of a Gaussian stationary process and $S(\nu)$ be its PSD. We assume $r(k) = \hat{r}_{XX}(k)\ \forall |k| < M$ and we want to maximize its entropy rate $H$ in equation (3.5) w.r.t.. $r(k)$ for $|k| \geq M$.

We thus want $\forall |k| \geq M$, $\frac{\partial H}{\partial r(k)} = 0$. However, since $S(\nu) = \sum_{k\in\mathbb{Z}} r(k)e^{-2\imath\pi\nu k}$, we get, by derivating (3.5),

$$\frac{\partial H}{\partial r(k)} = \frac{1}{2}\int_{-\frac{1}{2}}^{+\frac{1}{2}} \frac{\frac{\partial S(\nu)}{\partial r(k)}}{S(\nu)}d\nu = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \frac{1}{S(\nu)}e^{-2\imath\pi\nu k}d\nu.$$

Let $r_{YY}(k) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} \frac{1}{S(\nu)}e^{+2\imath\pi\nu k}d\nu$. Since $\frac{1}{S(\nu)} \geq 0$, $r_{YY}(k)$ is the ACF of a WSS process $Y_t$. However, we have just proved that $r_{YY}(k) = 0\ \forall |k| \geq M$. Then Theorem 16 proves that $Y_t$ is an MA($M-1$) process. Therefore there are coefficients $b_0 \dots b_{M-1}$ such that $\frac{1}{S(\nu)} = S_{YY}(\nu) = \sigma^2\left|\sum_{k=0}^{M-1} b_k e^{-2\imath\pi\nu k}\right|^2$. Finally, $S(\nu) = \frac{1}{S_{YY}(\nu)}$ is the PSD of an AR($M-1$) process. □

# 3 Linear prediction method for AR estimation

In this section, we remind the linear prediction approach for estimating AR models, the expression of Yule-Walker equations, we prove the stability of the causal AR filter obtained by solving these equations, and we remind the Levinson-Durbin algorithm that efficiently computes the solution.

## 3.1 Reminder: Estimation of an autoregressive process

↪ For more details on the content of this section, we refer the reader to [2, chap. 3].

The most popular way of estimating the parameters of a causal autoregressive process consists in predicting the value of the current sample $X_t$ as a linear combination of the $p$ previous samples:

$$\widehat{X}_t = \sum_{m=1}^{p} a_m X_{t-m}.$$

The *prediction coefficients* $a_1 \dots a_p$ are chosen so as to minimize the mean square error $\mathbb{E}[|Z_t|^2]$, where $Z_t$ denotes the prediction error $Z_t = X_t - \widehat{X}_t$. Solving this minimization problem is equivalent to decorrelating $Z_t$ from the

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

past samples $X_{t-k}$ (i.e. $\text{cov}(Z_t, X_{t-k}) = 0$) for $k \in \{1 \dots p\}$. This approach leads to the following set of linear equations: $\forall k \in \{1 \dots p\}$,

$$r_{XX}(k) = \sum_{j=1}^{p} a_j r_{XX}(k-j) \tag{3.6}$$

which can be solved to find the prediction coefficients $a_k$ given the ACF $r_{XX}(k)$, and

$$r_{XX}(0) = \sigma_Z^2 + \sum_{k=1}^{p} a_k r_{XX}(k), \tag{3.7}$$

which permits us to estimate $\sigma_Z^2$, since it implies $\sigma_Z^2 = r_{XX}(0) - \sum_{k=1}^{p} a_k r_{XX}(k)$.

These equations are known as the *Yule-Walker equations*. They can also be formulated in matrix form: we first estimate the covariance matrix $\boldsymbol{R}_{XX} = \mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^H]$, which can be defined indifferently from vector $\boldsymbol{X}_t = [X_t, X_{t+1} \dots X_{t+p-1}]^\top$ (forward data vector) or from vector $\boldsymbol{X}_t = [\overline{X_t}, \overline{X_{t-1}} \dots \overline{X_{t-p+1}}]^\top$ (backward data vector), leading to the same expression:

$$\boldsymbol{R}_{XX} = \begin{bmatrix} r_{XX}(0) & r_{XX}(-1) & \dots & r_{XX}(-(p-1)) \\ r_{XX}(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{XX}(-1) \\ r_{XX}(p-1) & \dots & r_{XX}(1) & r_{XX}(0) \end{bmatrix}. \tag{3.8}$$

The Yule-Walker equations (3.6) can then be rewritten in matrix-vector form:

$$\boldsymbol{R}_{XX} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_{XX}(1) \\ r_{XX}(2) \\ \vdots \\ r_{XX}(p) \end{bmatrix}. \tag{3.9}$$

In Section 3.2, we will show that the causal AR filter of transfer function $\frac{1}{1-\sum_{m=1}^{p} a_m z^{-m}}$ estimated by solving this system of equations is always stable. Then in Section 3.3, we will remind the fast Levinson-Durbin algorithm, which makes it possible to solve the Yule-Walker equations (3.9) and (3.7) in only $O(p^2)$ arithmetical operations instead of $O(p^3)$.

## 3.2 Stability of the AR filter solution of Yule-Walker equations

In this section, we prove that the causal AR filter estimated by solving Yule-Walker equations is always stable [15]. From now on, let $\boldsymbol{R}_n$ denote the $n \times n$ matrix defined as $\boldsymbol{R}_{XX}$ in equation (3.8), but with $p$ replaced by an arbitrary natural number $n$. Then the Yule-Walker equations (3.9) and (3.7) can be regrouped into a single equation:

$$\boldsymbol{R}_{n+1} \begin{pmatrix} 1 \\ -a_1 \\ \vdots \\ -a_n \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.10}$$

**Proposition 23.** *The matrix $\boldsymbol{R}_{n+1}$ in equation* (3.10) *is positive definite if and only if the transfer function $A(z) = 1 - \sum_{n=1}^{p} a_n z^{-n}$ is non-zero $\forall |z| \geq 1$.*

Proposition 23 proves that the causal AR filter estimated by solving Yule-Walker equations is always stable when $\boldsymbol{R}_n$ is the true covariance matrix of the AR process with $n = p$, since this matrix is positive definite.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

However, in practice the covariance matrix $\boldsymbol{R}_n$ has to be estimated. Fortunately, the estimate $\hat{\boldsymbol{R}}_n$ obtained by replacing the exact ACF $r_{XX}$ by its biased estimate $\hat{r}_{XX}$ introduced in equation (2.2) page 12 is always positive definite, provided that $[X_1 \dots X_n]^\top \neq \boldsymbol{0}$. Indeed, $\hat{\boldsymbol{R}}_n$ can be factorized as $\hat{\boldsymbol{R}}_n = \frac{1}{N}\boldsymbol{X}^H\boldsymbol{X}$ where $\boldsymbol{X}$ is the following $(N+n-1)\times n$ dimensional matrix:

$$\boldsymbol{X} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ X_2 & X_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ X_N & & \ddots & X_1 \\ 0 & \ddots & & X_2 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & X_N \end{bmatrix},$$

which proves that $\hat{\boldsymbol{R}}_n$ is positive semidefinite. In addition, assuming $[X_1 \dots X_n]^\top \neq \boldsymbol{0}$, matrix $\boldsymbol{X}$ has full rank, therefore $\hat{\boldsymbol{R}}_n$ is positive definite. Consequently, Proposition 23 proves that the causal AR filter estimated by solving Yule-Walker equations is always stable, even when matrix $\boldsymbol{R}_n$ is replaced by its biased estimate.

The following proof of Proposition 23 is provided for the interested reader only; it is outside the scope of the teaching unit's program.

*Proof of Proposition 23.* Let us introduce the following companion matrix:

$$\boldsymbol{A}_c = \begin{bmatrix} a_1 & 1 & 0 & \dots & 0 \\ a_2 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ a_{n-1} & 0 & \dots & 0 & 1 \\ a_n & 0 & \dots & \dots & 0 \end{bmatrix}. \tag{3.11}$$

The roots of $A(z)$ are equal to the eigenvalues of $\boldsymbol{A}_c$. Let $\lambda$ denote an arbitrary eigenvalue of $\boldsymbol{A}_c$, and $\boldsymbol{u} = [u_1 \dots u_n]^\top \neq \boldsymbol{0}$ the associated eigenvector

$$\boldsymbol{A}_c\boldsymbol{u} = \lambda\boldsymbol{u}. \tag{3.12}$$

It is easy to see from (3.11) and (3.12) that $\boldsymbol{u} \neq \boldsymbol{0}$ implies $u_1 \neq 0$. Thus, we can normalize $\boldsymbol{u}$ such that $u_1 = 1$. Then from (3.12), we get

$$\begin{cases} u_1 &= 1 \\ u_2 + a_1 &= \lambda u_1 \\ \vdots &= \vdots \\ u_n + a_{n-1} &= \lambda u_{n-1} \\ a_n &= \lambda u_n \end{cases}$$

which can be written more compactly as

$$\begin{bmatrix} \boldsymbol{u} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -\boldsymbol{a} \end{bmatrix} + \lambda \begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix} \tag{3.13}$$

where $\boldsymbol{a} = [a_1 \dots a_n]^\top$. It then follows from (3.10), (3.13), and the Toeplitz structure of $\boldsymbol{R}_{n+1}$ that

$$\begin{aligned} \boldsymbol{u}^H\boldsymbol{R}_n\boldsymbol{u} &= \begin{bmatrix} \boldsymbol{u}^H & 0 \end{bmatrix} \boldsymbol{R}_{n+1} \begin{bmatrix} \boldsymbol{u} \\ 0 \end{bmatrix} \\ &= \left( \begin{bmatrix} 1 & -\boldsymbol{a}^\top \end{bmatrix} + \overline{\lambda} \begin{bmatrix} 0 & \boldsymbol{u}^H \end{bmatrix} \right) \boldsymbol{R}_{n+1} \left( \begin{bmatrix} 1 \\ -\boldsymbol{a} \end{bmatrix} + \lambda \begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix} \right) \\ &= \sigma^2 + |\lambda|^2 \boldsymbol{u}^H\boldsymbol{R}_n\boldsymbol{u}. \end{aligned}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Thus, we get

$$|\lambda|^2 = 1 - \frac{\sigma^2}{\boldsymbol{u}^H \boldsymbol{R}_n \boldsymbol{u}}. \tag{3.14}$$

Finally, the equivalence in Proposition 23 follows immediately from (3.14), and the proof is concluded.

□

## 3.3 Reminder: Levinson-Durbin Algorithm

The Levinson-Durbin Algorithm solves equation (3.10) recursively in the order $n$. The only requirement is that the covariance matrix be positive definite, Hermitian, and Toeplitz. Thus, the algorithm applies equally well to the Yule-Walker AR estimator, in which the exact ACF coefficients are replaced by their biased estimates. Hence, to cover both cases jointly, we will use the common notation $\rho_k$ to denote either $r_{XX}(k)$ or $\hat{r}_{XX}(k)$. With this convention, we have

$$\boldsymbol{R}_{n+1} = \begin{bmatrix} \rho_0 & \rho_{-1} & \cdots & \rho_{-n} \\ \rho_1 & \rho_0 & & \vdots \\ \vdots & & \ddots & \rho_{-1} \\ \rho_n & \cdots & \rho_1 & \rho_0 \end{bmatrix} = \begin{bmatrix} \rho_0 & \overline{\rho_1} & \cdots & \overline{\rho_n} \\ \rho_1 & \rho_0 & & \vdots \\ \vdots & & \ddots & \overline{\rho_1} \\ \rho_n & \cdots & \rho_1 & \rho_0 \end{bmatrix}.$$

The following notational convention will also be frequently used in this section: for a vector $\boldsymbol{x} = [x_1 \dots x_n]^T$, we define $\tilde{\boldsymbol{x}} = [\overline{x_n} \dots \overline{x_1}]^T$. An important property of any Hermitian Toeplitz matrix $\boldsymbol{R}$ is that

$$\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x} \quad \Rightarrow \quad \tilde{\boldsymbol{y}} = \boldsymbol{R}\tilde{\boldsymbol{x}}. \tag{3.15}$$

The result (3.15) follows from the following calculation, with the change of variable $k = n - p + 1$:

$$\begin{aligned} \tilde{y}_i = \overline{y_{n-i+1}} &= \sum_{k=1}^{n} \overline{R_{n-i+1,k}}\,\overline{x_k} \\ &= \sum_{k=1}^{n} \overline{\rho_{n-i+1-k}}\,\overline{x_k} = \sum_{p=1}^{n} \overline{\rho_{p-i} x_{n-p+1}} = \sum_{p=1}^{n} R_{i,p}\tilde{x}_p \\ &= (\boldsymbol{R}\tilde{\boldsymbol{x}})_i \end{aligned}$$

where $R_{i,j}$ denotes the $(i, j)$ th element of the matrix $\boldsymbol{R}$. So the basic idea of the Levinson-Durbin algorithm is to solve (3.10) recursively in $n$, starting from the solution for $n = 1$ (which is easily determined). By using (3.10) and the nested structure of the covariance matrix, we can write

$$\boldsymbol{R}_{n+2} \begin{bmatrix} 1 \\ -\boldsymbol{a}_n \\ \hline 0 \end{bmatrix} = \left[\begin{array}{ccc|c} & \boldsymbol{R}_{n+1} & & \overline{\rho_{n+1}} \\ & & & \tilde{\boldsymbol{r}}_n \\ \hline \rho_{n+1} & \tilde{\boldsymbol{r}}_n^H & & \rho_0 \end{array}\right] \left[\begin{array}{c} 1 \\ -\boldsymbol{a}_n \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} \sigma_n^2 \\ \boldsymbol{0} \\ \hline \alpha_n \end{array}\right] \tag{3.16}$$

where

$$\begin{aligned} \boldsymbol{r}_n &= [\rho_1 \dots \rho_n]^T, \\ \alpha_n &= \rho_{n+1} - \tilde{\boldsymbol{r}}_n^H \boldsymbol{a}_n. \end{aligned}$$

Equation (3.16) would be the counterpart of (3.10) when $n$ is increased by one, if $\alpha_n$ in (3.16) could be nulled. To do so, let $k_{n+1} = -\alpha_n/\sigma_n^2$. It follows from (3.15) and (3.16) that

$$\begin{aligned} \boldsymbol{R}_{n+2} \left\{ \begin{bmatrix} 1 \\ -\boldsymbol{a}_n \\ 0 \end{bmatrix} + k_{n+1} \begin{bmatrix} 0 \\ -\tilde{\boldsymbol{a}}_n \\ 1 \end{bmatrix} \right\} &= \begin{bmatrix} \sigma_n^2 \\ \boldsymbol{0} \\ \alpha_n \end{bmatrix} + k_{n+1} \begin{bmatrix} \overline{\alpha_n} \\ \boldsymbol{0} \\ \sigma_n^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_n^2 + k_{n+1}\overline{\alpha_n} \\ \boldsymbol{0} \end{bmatrix} \end{aligned} \tag{3.17}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

which has the same structure as

$$\boldsymbol{R}_{n+2}\begin{bmatrix} 1 \\ -\boldsymbol{a}_{n+1} \end{bmatrix} = \begin{bmatrix} \sigma^2_{n+1} \\ \boldsymbol{0} \end{bmatrix}. \tag{3.18}$$

Comparing (3.17) and (3.18) and making use of the fact that the solution to (3.10) is unique for any $n$, we reach the conclusion that

$$\boldsymbol{a}_{n+1} = \begin{bmatrix} \boldsymbol{a}_n \\ 0 \end{bmatrix} + k_{n+1}\begin{bmatrix} \tilde{\boldsymbol{a}}_n \\ -1 \end{bmatrix} \tag{3.19}$$

and

$$\sigma^2_{n+1} = \sigma^2_n\left(1 - |k_{n+1}|^2\right) \tag{3.20}$$

constitute the solution to (3.10) for order $(n+1)$.

Equations (3.19) and (3.20) form the core of the Levinson-Durbin algorithm. The initialization of these recursive-in-$n$ equations is straightforward. Algorithm 1 below summarizes the Levinson-Durbin method in a form that should be convenient for machine coding. The Levinson-Durbin algorithm has many interesting properties and uses for which we refer to [16, 1, 17]. The coefficients $k_n$ in the Levinson-Durbin algorithm are often called the *reflection coefficients*; $-k_n$ are also called the *partial correlation* (PARCOR) coefficients.

**Algorithm 1** The Levinson-Durbin Algorithm

Initialization:

$a_1 = \rho_1/\rho_0 = -k_1$ [1 flop ]

$\sigma^2_1 = \rho_0 - |\rho_1|^2/\rho_0$ [1 flop ]

For $n = 1, \ldots, n_{\max}$, do:

$$k_{n+1} = -\frac{\rho_{n+1} - \tilde{\boldsymbol{r}}_n^H \boldsymbol{a}_n}{\sigma^2_n} \qquad [n+1 \text{ flops }]$$

$$\sigma^2_{n+1} = \sigma^2_n\left(1 - |k_{n+1}|^2\right) \qquad [2 \text{ flops }]$$

$$\boldsymbol{a}_{n+1} = \begin{bmatrix} \boldsymbol{a}_n \\ 0 \end{bmatrix} + k_{n+1}\begin{bmatrix} \tilde{\boldsymbol{a}}_n \\ -1 \end{bmatrix} \qquad [n \text{ flops }]$$

It can be seen from Algorithm 1 that the Levinson-Durbin method requires on the order of $2n$ flops to compute $\left\{\boldsymbol{a}_{n+1}, \sigma^2_{n+1}\right\}$ from $\left\{\boldsymbol{a}_n, \sigma^2_n\right\}$. Hence a total of about $n^2_{\max}$ flops is needed to compute all the solutions to the Yule-Walker system of equations, from $n = 1$ to $n = n_{\max}$.

# 4 Reminder: ARMA processes

In this section, we remind a few known properties of ARMA models.

## 4.1 ARMA$(p, q)$ process

**Definition 14** (ARMA$(p, q$ process)**.** *Consider the recursive equation:*

$$X_t - a_1 X_{t-1} - \ldots - a_p X_{t-p} = Z_t + b_1 Z_{t-1} + \ldots + b_q Z_{t-q}, \tag{3.21}$$

*where $Z_t \sim \mathrm{WN}(0, \sigma^2)$ and $a_n, b_n \in \mathbb{C}$. Let $A(z) = 1 - a_1 z^{-1} - \ldots - a_p z^{-p}$ and $B(z) = 1 + b_1 z^{-1} + \ldots + b_q z^{-q}$. We assume that $A(z)$ and $B(z)$ do not have common zeros.*

*Then equation* (3.21) *admits a WSS solution if and only if $A(z) \neq 0\ \forall |z| = 1$. This solution is unique and its expression is $X_t = \sum_{n=-\infty}^{+\infty} h_n Z_{t-n}$, where the $h_n$ are given by the coefficients of the expansion $\frac{B(z)}{A(z)} = \sum_{n=-\infty}^{+\infty} h_n z^{-n}$, converging in the ring $\{z \in \mathbb{C}, \delta_1 < |z| < \delta_2\}$, where $\delta_1 < 1$ and $\delta_2 > 1$ are defined by $\delta_1 = \max\{z \in \mathbb{C}, |z| < 1, A(z) = 0\}$ and $\delta_2 = \min\{z \in \mathbb{C}; |z| > 1; A(z) = 0\}$.*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Proposition 24** (Spectral density of an ARMA$(p,q)$ process)**.** *Let $X_t$ be an ARMA$(p,q)$ process, i.e. the WSS solution of equation* (3.21)*, where $B(z)$ and $A(z)$ are polynomials of degree q and p which do not have common zeros and $A(z) \neq 0$ $\forall |z| = 1$. Then $X_t$ is centered and has a spectral density whose expression is:*

$$S_{XX}(\nu) = \sigma^2 \frac{\left|1 + \sum_{n=1}^{q} b_n e^{-2\imath\pi\nu n}\right|^2}{\left|1 - \sum_{n=1}^{p} a_n e^{-2\imath\pi\nu n}\right|^2}.$$

## 4.2 Representations of an ARMA$(p,q)$ process

**Definition 15.** *Let $X_t$ be an ARMA$(p,q)$ process solution of equation* (3.21)*. Then $X_t$ admits a linear representation $X_t = \sum_{n=-\infty}^{+\infty} h_n Z_{t-n}$ for a well chosen sequence $h_n \in l^1(\mathbb{Z})$. We say that the ARMA$(p,q)$ representation is*

- causal *if the filter $h_n$ is causal ($A(z) \neq 0$ $\forall |z| \geq 1$);*
- invertible *if the filter $h_n$ is invertible and if its inverse is causal ($B(z) \neq 0$ $\forall |z| \geq 1$);*
- canonical *if it is causal and invertible (i.e. the filter $h_n$ is minimum phase).*

**Theorem 25** (Canonical representation)**.** *Let $X_t$ be an ARMA$(p,q)$ process solution of equation* (3.21)*. We assume that $A(z) \neq 0$ and $B(z) \neq 0$ $\forall |z| = 1$. Then $X_t$ admits a canonical representation.*

*Proof.* Any recursive filter of transfer function $\frac{B(z)}{A(z)}$ such that that $A(z) \neq 0$ and $B(z) \neq 0$ $\forall |z| = 1$ can be factorized as the product of a minimum phase filter and an all-pass filter. □

## 4.3 Covariances of a causal ARMA process

In this section, we address the problem of calculating the ACF of a causal ARMA process given its parameters. The inverse problem, which consists in estimating the parameters of a causal ARMA process given its ACF, will be addressed in Section 5.

To calculate the ACF of a causal ARMA process, a first approach consists in using the expression $r_{XX}(k) = \sigma^2 \sum_{n=0}^{+\infty} h_{n+k}\overline{h}_n$ where $h_n$ is determined recursively from $H(z)A(z) = B(z)$, by identification of the term in $z^{-n}$. For the first terms, we find:

$$\begin{aligned} h_0 &= 1 \\ h_1 &= b_1 + h_0 a_1 \\ h_2 &= b_2 + h_0 a_2 + h_1 a_1 \end{aligned}$$

Another approach is based on a recursion formula, verified by the ACF of an ARMA$(p,q)$ process, which is obtained by multiplying with $\overline{X}_{t-n}$ for $n > q$ the two members of equation (3.21), and by taking the mathematical expectation. This recursion formula will be given in Section 5.1. It makes it possible to compute $r_{XX}(n)$ recursively.

# 5 Durbin method for ARMA estimation

Let $X_t$ be a causal ARMA$(p,q)$ process as defined by the recursive equation (3.21), where $b_0 = 1$ and $Z_t \sim \text{WN}(0, \sigma_Z^2)$. In this section, we aim to estimate its parameters given an estimate of its ACF.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

## 5.1 Estimation of the AR part (modified Yule-Walker Method)

By multiplying the equation (3.21) with $\overline{X_{t-n}}$ for $n > q$, and by applying the mathematical expectation, we get

$$r_{XX}(n) - \sum_{k=1}^{p} a_k r_{XX}(n-k) = \sum_{k=0}^{q} b_k \mathbb{E}[\overline{X_{t-n}} Z_{t-k}].$$

However, because the processus $X_t$ is causal and $n > q$, the second member of this equation is zero.

By stacking the $p$ equations $r_{XX}(n) - \sum_{k=1}^{p} a_k r_{XX}(n-k) = 0$ for all $n \in \{q+1 \dots q+p\}$, we finally get the set of *modified Yule-Walker equations*:

$$\begin{bmatrix} r_{XX}(q+1) \\ \vdots \\ r_{XX}(q+p) \end{bmatrix} = \begin{bmatrix} r_{XX}(q) & r_{XX}(q-1) & \dots & r_{XX}(q-p+1) \\ r_{XX}(q+1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{XX}(q-1) \\ r_{XX}(q+p-1) & \dots & r_{XX}(q+1) & r_{XX}(q) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix},$$

which can be solved to find estimates $\hat{a}_1 \dots \hat{a}_p$ of the prediction coefficients $a_1 \dots a_p$.

We now need to estimate the MA coefficients $b_1 \dots b_q$.

## 5.2 Estimation of the PSD: first approach

Consider the WSS process

$$Y_t = X_t - a_1 X_{t-1} - \dots - a_p X_{t-p},$$

whose samples can be directly computed from those of $X_t$ and from the estimated prediction coefficients $a_1 \dots a_p$. We note that

$$S_{XX}(\nu) = \frac{S_{YY}(\nu)}{|A(\nu)|^2} \tag{3.22}$$

where $A(\nu) = 1 - \sum_{m=1}^{p} a_m z^{-m}$. Equation (3.21) shows that $Y_t$ can also be expressed as

$$Y_t = b_0 Z_t + b_1 Z_{t-1} + \dots + b_q Z_{t-q},$$

therefore $Y_t$ is actually an MA($q$) process. Its ACF $r_{YY}(k)$ for $k \in [-q \dots q]$ can be directly estimated from the samples $Y_t$. Then its PSD can be estimated by applying a DTFT to the estimated sequence $\hat{r}_{YY}(k)$:

$$\hat{S}_{YY}(\nu) = \sum_{k=-q}^{q} \hat{r}_{YY}(k) e^{-2\imath\pi\nu k}. \tag{3.23}$$

By substituting (3.23) into (3.22), we get a first ARMA PSD estimate:

$$\hat{S}_{XX}(\nu) = \frac{\sum_{k=-q}^{q} \hat{r}_{YY}(k) e^{-2\imath\pi\nu k}}{\left|\hat{A}(\nu)\right|^2}.$$

where $\hat{A}(\nu) = 1 - \sum_{m=1}^{p} \hat{a}_m z^{-m}$. The drawbacks of this PSD estimate are that the numerator is not necessarily non-negative, and that we do not get estimates of the MA coefficients $b_1 \dots b_q$.

## 5.3 Estimation of the MA part: Durbin method

In order to get a non-negative estimate of the ARMA PSD, we will now introduce the *Durbin method*.

First, we consider the estimate $\hat{r}_{YY}(k)$ of the ACF of the MA process $Y_t$. We want to find MA coefficients $\hat{b}_1 \dots \hat{b}_q, \hat{\sigma}_Z^2$ such that $\hat{S}_{YY}(\nu) = \hat{\sigma}_Z^2 |\hat{B}(\nu)|^2$ with $\hat{B}(z) = 1 + \sum_{n=1}^{q} \hat{b}_n z^{-q}$.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

To do so, we first solve the Yule-Walker equations to find an AR($L$) model that fits the sequence $\hat{r}_{YY}(k)$ for $k = 0 \dots L \gg q$. We thus find prediction coefficients $\hat{a}_{1,L}, \dots, \hat{a}_{L,L}$ such that

$$\hat{S}_{YY}(\nu) = \frac{\sigma_1^2}{|\hat{A}_L(\nu)|^2} \tag{3.24}$$

with $\hat{A}_L(\nu) = 1 - \sum_{m=1}^{L} \hat{a}_{m,L} z^{-m}$. Then we consider the sequence $\hat{r}_L(k)$ such that

$$|\hat{A}_L(\nu)|^2 = \sum_{k=-L}^{L} \hat{r}_L(k) e^{-2\imath\pi\nu k}.$$

We now solve another set of Yule-Walker equations to find an AR($q$) model that fits the sequence $\hat{r}_L(k)$. We thus find prediction coefficients $[\hat{b}_1, \dots, \hat{b}_q]$ such that

$$|\hat{A}_L(\nu)|^2 = \frac{\sigma_2^2}{|\hat{B}(\nu)|^2} \tag{3.25}$$

with $\hat{B}(z) = 1 + \sum_{n=1}^{q} \hat{b}_n z^{-q}$. Substituting (3.25) into (3.24), we get

$$\hat{S}_{YY}(\nu) = \frac{\sigma_1^2}{\sigma_2^2} |\hat{B}(\nu)|^2.$$

Finally, from equation (3.22), we get the following ARMA PSD estimate:

$$\hat{S}_{XX}(\nu) = \frac{\hat{S}_{YY}(\nu)}{|\hat{A}(\nu)|^2} = \hat{\sigma}_Z^2 \frac{|\hat{B}(\nu)|^2}{|\hat{A}(\nu)|^2}$$

with $\hat{\sigma}_Z^2 = \frac{\sigma_1^2}{\sigma_2^2}$, which is always non-negative.

# Chapter 4

# Filter bank methods

In Chapter 2, we have seen that the periodogram and correlogram methods provide reasonably high resolution for sufficiently long data lengths, but are poor spectral estimators because their variance is high and does not decrease with increasing data length. In this chapter, we provide an interpretation of the periodogram and correlogram methods as a power estimate based on a single sample of a filtered version of the signal under study; it is thus not surprising that the periodogram or correlogram variance is large. This chapter makes the transition from the periodogram approach presented in Chapter 2 and the linear prediction techniques presented in Chapter 3, to the methods dedicated to the estimation of line spectra, that will be introduced in Chapter 5.

This chapter is structured as follows: in Section 1, we first present the filter bank interpretation of the periodogram. Then in Section 2, we introduce Capon's spectral estimator and a variant by Lagunas, before analyzing the statistical properties of Capon's method.

↪ For more details on the content of this chapter, we refer the reader to [2, chap. 5].

## 1 The periodogram as a filter bank

From Definition 8 in Chapter 2 page 13, we know that the periodogram $\hat{S}_{XX}(\nu)$ of a centered WSS random process $X_t$ is defined as $\hat{S}_{XX}(\nu) = \frac{1}{N}\left|\sum_{t=0}^{N-1} X_t e^{-2\imath\pi\nu t}\right|^2$. The following result provides us with a filter bank interpretation of the periodogram.

**Proposition 26** (Filter bank interpretation of the periodogram)**.** *Let $X_t$ be a centered WSS process. Its periodogram as introduced in Definition 8 can also be expressed as*

$$\hat{S}_{XX}(\nu) = N\left|\sum_{k\in\mathbb{Z}} w_k^\nu X_{N-1-k}\right|^2 = \frac{|Y_{N-1}|^2}{1/N}, \tag{4.1}$$

*where $Y_t = w^\nu * X_t$, $w_k^\nu = 0$ if $k \notin [0\dots N-1]$, and $w_k^\nu = \frac{1}{N} e^{2\imath\pi\nu k}$ if $k \in [0\dots N-1]$.*

*Proof of Proposition 26.* We have $\hat{S}_{XX}(\nu) = \frac{1}{N}\left|\sum_{t=0}^{N-1} X_t e^{-2\imath\pi\nu t}\right|^2$. Let $k = N-1-t$; we then get $\hat{S}_{XX}(\nu) = N\left|\frac{1}{N}\sum_{k=0}^{N-1} e^{2\imath\pi\nu k} X_{N-1-k}\right|^2$. However, we note that $Y_{N-1} = \sum_{k\in\mathbb{Z}} w_k^\nu X_{N-1-k} = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\imath\pi\nu k} X_{N-1-k}$. Therefore we finally get equation (4.1). □

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

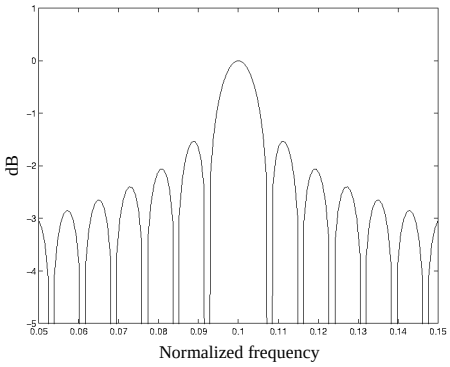



Figure 4.1: Frequency response $\xi \mapsto |W^\nu(\xi)|$ with $\nu = 0.1$

Proposition 26 shows that the periodogram is equivalent to filtering the signal $X_t$ by a FIR filter, and computing the power of the $(N-1)^{th}$ output sample. We also note that the magnitude of the DTFT of $w^\nu$ is $|W^\nu(\xi)| = \left|\sum_{k\in\mathbb{Z}} w_k^\nu e^{-2\imath\pi\xi k}\right| = \frac{1}{N}\left|\frac{\sin(\pi(\xi-\nu)N)}{\sin(\pi(\xi-\nu))}\right|$. The resulting frequency response $W^\nu$ has a main lobe with a small width $\frac{2}{N}$, but side lobes with high amplitudes (see Figure 4.1).

# 2 Capon's method

This result emphasizes a limitation of the periodogram approach: $Y_t$ has a non-zero power on a large frequency band around $\nu$, so that the estimation of $\hat{S}_{XX}(\nu)$ is biased. In contrast, Capon's method that will be introduced in this section consists in determining a filter $w^\nu$ such that $W^\nu(\nu) = 1$, which minimizes the power of the output signal at frequencies other than $\nu$, so as to minimize the spectral leakage.

**Lemma 27.** *Let $X_t$ be a centered WSS process. Let $N \in \mathbb{N}^*$ and $\boldsymbol{X} = [\overline{X_{N-1}} \dots \overline{X_0}]^\top$; we assume that the covariance matrix $\boldsymbol{R}_{XX} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^H]$ is positive definite, thus non-singular.*

*Let $Y = w^\nu * X$, where $w^\nu$ is a FIR filter of support $[0 \dots N-1]$. Finally, let $\boldsymbol{e}(\xi) = [1, e^{2\imath\pi\xi} \dots e^{2\imath\pi\xi(N-1)}]^\top$ and $\boldsymbol{w}^\nu = [w_0^\nu \dots w_{N-1}^\nu]^\top$, and suppose that $W^\nu(\nu) = \boldsymbol{e}(\nu)^H \boldsymbol{w}^\nu = 1$. Then*

$$\mathbb{E}[|Y_{N-1}|^2] = (\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu)^H \boldsymbol{R}_{XX} (\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu) + \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)},$$

*where*

$$\boldsymbol{w}_{\text{opt}}^\nu = \frac{\boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu)}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}. \tag{4.2}$$

*Proof of Lemma 27.* We have $Y_{N-1} = \sum_{k\in\mathbb{Z}} w_k^\nu X_{N-1-k} = \boldsymbol{X}^H \boldsymbol{w}^\nu$. Therefore $\mathbb{E}[|Y_{N-1}|^2] = \mathbb{E}[\boldsymbol{w}^{\nu H} \boldsymbol{X}\boldsymbol{X}^H \boldsymbol{w}^\nu] = \boldsymbol{w}^{\nu H} \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^H] \boldsymbol{w}^\nu = \boldsymbol{w}^{\nu H} \boldsymbol{R}_{XX} \boldsymbol{w}^\nu$. Besides, we have $W^\nu(\xi) = \sum_{k=0}^{N-1} w_k^\nu e^{-2\imath\pi\nu k} = \boldsymbol{e}(\xi)^H \boldsymbol{w}^\nu$.

In other respects, since $\boldsymbol{w}_{\text{opt}}^\nu = \frac{\boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu)}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu)}$, we have

$$(\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu)^H \boldsymbol{R}_{XX} (\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu) + \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} = \boldsymbol{w}^{\nu H} \boldsymbol{R}_{XX} \boldsymbol{w}^\nu - \frac{\boldsymbol{e}(\nu)^H \boldsymbol{w}^\nu}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} - \frac{\boldsymbol{w}^{\nu H} \boldsymbol{e}(\nu)}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} + \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} + \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}.$$

Finally, since $\boldsymbol{e}(\nu)^H \boldsymbol{w}^\nu = 1$, we get the simplification

$$(\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu)^H \boldsymbol{R}_{XX} (\boldsymbol{w}^\nu - \boldsymbol{w}_{\text{opt}}^\nu) + \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} = \boldsymbol{w}^{\nu H} \boldsymbol{R}_{XX} \boldsymbol{w}^\nu = \mathbb{E}[|Y_{N-1}|^2].$$

□

We can now introduce the *Minimum Variance Distortionless Response* (MVDR) filter:

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Definition 16** (MVDR filter)**.** *Let $X_t$ be a centered WSS process. Let $Y = w^\nu * X$, where $w^\nu$ is a FIR filter of support $[0 \dots N-1]$. The MVDR filter is defined as the unique filter $w^\nu$ which minimizes $\mathbb{E}[|Y_{N-1}|^2]$ (*Minimum Variance*), subject to $W^\nu(\nu) = 1$ (*Distortionless Response*).*

The following proposition directly follows from Lemma 27:

**Proposition 28.** *With the same notation as in Lemma 27, the MVDR filter introduced in Definition 16 is $\boldsymbol{w}^\nu = \boldsymbol{w}^\nu_{\text{opt}}$. The power at the output is then $\mathbb{E}[|Y_{N-1}|^2] = \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}$.*

We can now introduce Capon's spectral estimator, which just consists in replacing filter $w^\nu$ in equation (4.1) by the MVDR filter:

**Definition 17** (Capon's spectral estimator)**.** *Capon's spectral estimator is*

$$\hat{S}_{\text{CAP},XX}(\nu) = \frac{\mathbb{E}[|Y_{N-1}|^2]}{1/N} = \frac{{\boldsymbol{w}^\nu_{\text{opt}}}^H \boldsymbol{R}_{XX} \boldsymbol{w}^\nu_{\text{opt}}}{1/N} = \frac{N}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}. \tag{4.3}$$

Note that in practice, $\boldsymbol{R}_{XX}$ is unknown and has to be estimated.
Also note that in the particular case of a white noise of variance $\sigma_X^2$, we have $\boldsymbol{R}_{XX} = \sigma_X^2 \boldsymbol{I}_N$, therefore

$$\boldsymbol{w}^\nu_{\text{opt}} = \frac{\boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)} = \frac{1}{N} \boldsymbol{e}(\nu).$$

Consequently, $w_k^\nu = \frac{1}{N} e^{2\imath\pi\nu k}\ \forall k \in [0 \dots N-1]$ is the same filter as the one involved in the periodogram (see Proposition 26).

## 2.1 Example: AR(1) process

We now consider the example of a causal AR(1) process: let $X_t = a_1 X_{t-1} + Z_t$ where $a_1 \in ]0, 1[$ and $Z_t \sim \text{WN}(0, \sigma_Z^2)$. The PSD of this WSS process is $S_{XX}(\nu) = \frac{\sigma_Z^2}{|1 - a_1 e^{-2\imath\pi\nu}|^2}$ (represented in Figure 4.2).

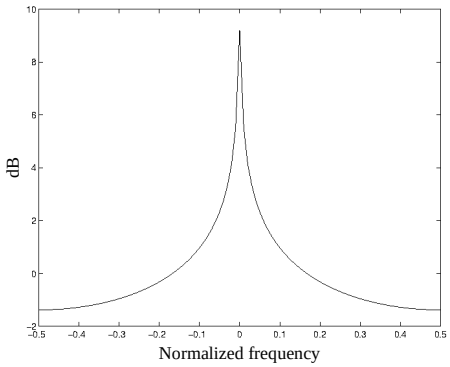



Figure 4.2: PSD of the AR process: $\nu \mapsto |S_{XX}(\nu)|$ with $a_1 = 0.99$ and $\sigma_Z^2 = 1$

The covariance matrix can be easily calculated:

$$\boldsymbol{R}_{XX} = \frac{\sigma_Z^2}{1 - a_1^2} \begin{bmatrix} 1 & a_1 & \dots & a_1^{N-1} \\ a_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_1^{N-1} & \dots & a_1 & 1 \end{bmatrix}.$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Its inverse matrix is

$$\boldsymbol{R}_{XX}^{-1} = \frac{1}{\sigma_Z^2} \begin{bmatrix} 1 & -a_1 & 0 & \dots & 0 \\ -a_1 & 1+a_1^2 & -a_1 & \ddots & \vdots \\ 0 & -a_1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+a_1^2 & -a_1 \\ 0 & \dots & 0 & -a_1 & 1 \end{bmatrix}$$

(this result can be easily proved by just checking that $\boldsymbol{R}_{XX}\boldsymbol{R}_{XX}^{-1} = \boldsymbol{I}$).

We thus get $\boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu) = \frac{|1-a_1 e^{-2\imath\pi\nu}|^2}{\sigma_Z^2}\,(\boldsymbol{e}(\nu) + \boldsymbol{v}(\nu))$ where $\boldsymbol{v}(\nu) = a_1\,[\frac{e^{-2\imath\pi\nu}}{1-a_1 e^{-2\imath\pi\nu}}, 0, \dots, 0, \frac{e^{+2\imath\pi N\nu}}{1-a_1 e^{+2\imath\pi\nu}}]^\top$.

Therefore the MVDR filter is $\boldsymbol{w}_{\text{opt}}^{\nu} = \frac{\boldsymbol{R}_{XX}^{-1}\,\boldsymbol{e}(\nu)}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu)} = \frac{\boldsymbol{e}(\nu)+\boldsymbol{v}(\nu)}{N+2a_1\frac{\cos(2\pi\nu)-a_1}{1-2a_1\cos(2\pi\nu)+a_1^2}}$.

Figure 4.3 shows that the shape of its frequency response depends on the targeted frequency $\nu$.

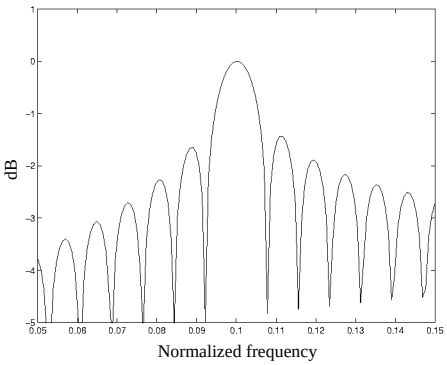



Figure 4.3: Frequency response $\xi \mapsto |W^{\nu}(\xi)|$ with $\nu = 0.1$ and $a_1 = 0.99$

## 2.2 Variant: Lagunas' method

In Proposition 28, we have proved that the MVDR filter leads to

$$\mathbb{E}[|Y_{N-1}|^2] = \frac{1}{\boldsymbol{e}(\nu)^H \boldsymbol{R}_{XX}^{-1}\,\boldsymbol{e}(\nu)}.$$

This is the power of $X_t$ in a passband centered on $\nu$. Then we can approximately determine the value of the PSD of $X_t$ at the passband's center frequency as

$$S_{XX}(\nu) \simeq \frac{\mathbb{E}\left[|Y_{N-1}|^2\right]}{\beta} = \frac{1}{\beta\,\boldsymbol{e}^H(\nu)\boldsymbol{R}_{XX}^{-1}\boldsymbol{e}(\nu)} \tag{4.4}$$

where $\beta$ denotes the frequency bandwidth of the filter $\boldsymbol{w}_{\text{opt}}^{\nu}$ given by (4.2). Note that since the bandpass filter (4.2) is data dependent, its bandwidth $\beta$ is not necessarily data independent, nor is it necessarily frequency independent. Hence, the division by $\beta$ in (4.4) may not represent a simple scaling of $\mathbb{E}\left[|Y_{N-1}|^2\right]$, but it may change the shape of this quantity as a function of $\nu$. There are various possibilities for determining the bandwidth $\beta$, depending on the degree of precision we are aiming for. The simplest possibility is to set

$$\beta = \frac{1}{N} \tag{4.5}$$

as in Capon's spectral estimator introduced in Definition 17. This choice is motivated by the time-bandwidth product formula $N\beta = 1$, which says that for a filter whose temporal aperture is equal to $N$, the bandwidth should roughly be given by $\frac{1}{N}$. By inserting (4.5) in (4.4), we retrieve (4.3).

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Note that if $X_t$ is white noise of variance $\sigma^2$, (4.3) takes the correct value: $S_{XX}(\nu) = \sigma^2$. In the general case, however, (4.5) gives only a rough indication of the filter's bandwidth.

An often more exact expression for $\beta$ can be obtained as follows [18]. The (equivalent) bandwidth of a bandpass filter can be defined as the support of the rectangle centered on $\nu$ (the filter's center frequency) that concentrates the whole energy in the filter's frequency response. According to this definition, $\beta$ can be assumed to satisfy:

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} |W(\xi)|^2 d\xi = |W(\nu)|^2 \beta \tag{4.6}$$

Since in the present case $W(\nu) = 1$ (*cf.* Definition 16), we obtain from (4.6):

$$\beta = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| \boldsymbol{w}^H \boldsymbol{e}(\xi) \right|^2 d\xi = \boldsymbol{w}^H \left[ \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{e}(\xi) \boldsymbol{e}^H(\xi) d\xi \right] \boldsymbol{w} \tag{4.7}$$

The $(k, p)$ element of the central matrix in the above quadratic form is given by

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-2\imath\pi\xi(k-p)} d\xi = \delta_{k,p}$$

With this observation and (4.2), (4.7) leads to

$$\beta = \boldsymbol{w}^H \boldsymbol{w} = \frac{\boldsymbol{e}^H(\nu) \boldsymbol{R}_{XX}^{-2} \boldsymbol{e}(\nu)}{\left[ \boldsymbol{e}^H(\nu) \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu) \right]^2}. \tag{4.8}$$

Note that this expression of the bandwidth is both data and frequency dependent (as was remarked previously). Inserting (4.8) in (4.4) gives

$$S_{XX}(\nu) \simeq \frac{\boldsymbol{e}^H(\nu) \boldsymbol{R}_{XX}^{-1} \boldsymbol{e}(\nu)}{\boldsymbol{e}^H(\nu) \boldsymbol{R}_{XX}^{-2} \boldsymbol{e}(\nu)}. \tag{4.9}$$

In the derivations above, the true data covariance matrix $\boldsymbol{R}_{XX}$ has been assumed available. In order to turn the previous PSD formulas into practical spectral estimation algorithms, we must replace $\boldsymbol{R}_{XX}$ in these formulas by a sample estimate, for instance by

$$\hat{\boldsymbol{R}}_{XX} = \frac{1}{L-N+1} \sum_{t=N}^{L} \begin{bmatrix} \overline{X_t} \\ \vdots \\ \overline{X_{t-N+1}} \end{bmatrix} [X_t \dots X_{t-N+1}] \tag{4.10}$$

where $L$ is the total number of observed signal samples. Doing so, we obtain the following two spectral estimators corresponding to (4.3) and (4.9), respectively:

$$\text{Capon:} \quad \hat{S}_{\text{CAP},XX}(\nu) = \frac{N}{\boldsymbol{e}^H(\nu) \hat{\boldsymbol{R}}_{XX}^{-1} \boldsymbol{e}(\nu)}; \tag{4.11}$$

$$\text{Lagunas:} \quad \hat{S}_{\text{LAG},XX}(\nu) = \frac{\boldsymbol{e}^H(\nu) \hat{\boldsymbol{R}}_{XX}^{-1} \boldsymbol{e}(\nu)}{\boldsymbol{e}^H(\nu) \hat{\boldsymbol{R}}_{XX}^{-2} \boldsymbol{e}(\nu)}. \tag{4.12}$$

There is an implicit assumption in both (4.11) and (4.12) that $\hat{\boldsymbol{R}}_{XX}^{-1}$ exists. This assumption sets a limit on the maximum value that can be chosen for $N$ :

$$N - 1 < L/2 \tag{4.13}$$

(observe that $\text{rank}(\hat{\boldsymbol{R}}_{XX}) \leq L-N+1$, which is less than $\dim(\hat{\boldsymbol{R}}_{XX}) = N$ if (4.13) is violated). The inequality (4.13) is important since it sets a limit on the resolution achievable by the Capon method. Indeed, since the Capon method is based on a bandpass filter with impulse response's aperture equal to $N - 1$, we may expect its resolution threshold

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

to be on the order of $\frac{1}{N-1} > \frac{2}{L}$ (with the inequality following from (4.13)). As $N$ is decreased, we can expect the resolution of Capon method to become worse (*cf.* the previous discussion). On the other hand, the accuracy with which $\hat{\boldsymbol{R}}_{XX}$ is determined increases with decreasing $N$ (since more outer products are averaged in (4.10)). The main consequence of the increased accuracy of $\hat{\boldsymbol{R}}_{XX}$ is to statistically stabilize the spectral estimate (4.11) or (4.12). Hence, the choice of $N$ should be done with the ubiquitous trade-off between resolution and statistical accuracy in mind. It is interesting to note that for the Capon method both the filter design and power calculation stages are data dependent. The accuracy of both these stages may worsen if $N$ is chosen too large. In applications, the maximum value that can be chosen for $N$ might also be limited from considerations of computational complexity.

Empirical studies have shown that the ability of the Capon method to resolve fine details of a PSD, such as closely spaced peaks, is superior to the corresponding performance of the periodogram-based methods.

## 2.3 Statistical properties of Capon's method

Capon's estimator is known to have a better spectral resolution than the periodogram. Moreover, its variance is lower than that of autoregressive methods, but its spectral resolution is worse, as shown by the following proposition:

**Proposition 29.** *If the estimator $\hat{S}_{\mathrm{CAP},XX}(\nu)$ in equation* (4.11) *is calculated with an $N \times N$ matrix $\hat{\boldsymbol{R}}_{XX}^{-1}$, using the covariance matrix estimate* (4.10)*, then $\hat{S}_{\mathrm{CAP},XX}(\nu)$ is related to AR estimators through*

$$\frac{1}{\hat{S}_{\mathrm{CAP},XX}(\nu)} = \frac{1}{N}\sum_{k=0}^{N-1} \frac{1}{\hat{S}_{\mathrm{AR}(k),XX}(\nu)} \tag{4.14}$$

*where $\hat{S}_{\mathrm{AR}(k),XX}(\nu)$ denotes the AR estimator of order $k$, with covariance matrix defined as the $k \times k$ lower right block of $\hat{\boldsymbol{R}}_{XX}$.*

Indeed, relation (4.14) says that the inverse of the Capon spectral estimator can be obtained by averaging the inverse estimated AR spectra of orders from 0 to $N-1$. In view of the averaging operation in (4.14), it is not difficult to understand why the Capon estimator possesses less statistical variability than the AR estimator. Moreover, the fact that the Capon estimator has also been found to have worse resolution and bias properties than the AR spectral estimate should be due to the presence of low-order AR models in (4.14).

The following proof of Proposition 29 is provided for the interested reader only; it is outside the scope of the teaching unit's program.

*Proof of Proposition 29.* The AR method of spectral estimation has been described in Chapter 3. In particular, the AR method corresponding to the sample covariance matrix (4.10) is called the *Least Squares* (LS) method. Let us denote the matrix $\hat{\boldsymbol{R}}_{XX}$ in (4.10) by $\hat{\boldsymbol{R}}_N$ and its principal $k \times k$ lower right block by $\hat{\boldsymbol{R}}_k\,(k = 1, \ldots, N)$, as shown below:

$$\boxed{\hat{\boldsymbol{R}}_N \quad \boxed{\hat{\boldsymbol{R}}_k \quad \boxed{\hat{\boldsymbol{R}}_1}}} \tag{4.15}$$

With this notation, the coefficient vector $\boldsymbol{a}_k$ and the residual power $\sigma_k^2$ of the $k$-th order AR model fitted to the data $\{X_t\}$ are obtained as the solutions to the following matrix equation (*cf.* equation (3.10) page 27):

$$\hat{\boldsymbol{R}}_{k+1}\begin{bmatrix} 1 \\ -\hat{\boldsymbol{a}}_k \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_k^2 \\ 0 \end{bmatrix} \tag{4.16}$$

The nested structure of (4.15) along with equation (4.16) imply:

$$\hat{\boldsymbol{R}}_N \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ & 1 & \ddots & \vdots & \vdots \\ & & \ddots & 0 & \vdots \\ & & & 1 & 0 \\ -\hat{\boldsymbol{a}}_{N-1} & -\hat{\boldsymbol{a}}_{N-2} & & -\hat{\boldsymbol{a}}_1 & 1 \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_{N-1}^2 & x & \ldots & x \\ 0 & \hat{\sigma}_{N-2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & x \\ 0 & \cdots & 0 & \hat{\sigma}_0^2 \end{bmatrix} \tag{4.17}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

where " $x$ " stands for undetermined elements. Let

$$\hat{\boldsymbol{\mathcal{H}}} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ & 1 & \ddots & \vdots & \vdots \\ & & \ddots & 0 & \vdots \\ & & & 1 & 0 \\ -\hat{\boldsymbol{a}}_{N-1} & -\hat{\boldsymbol{a}}_{N-2} & & -\hat{\boldsymbol{a}}_1 & 1 \end{bmatrix}.$$

It follows from (4.17) that

$$\hat{\boldsymbol{\mathcal{H}}}^H \hat{\boldsymbol{R}}_N \hat{\boldsymbol{\mathcal{H}}} = \begin{bmatrix} \hat{\sigma}^2_{N-1} & x & \dots & x \\ 0 & \hat{\sigma}^2_{N-2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & x \\ 0 & \dots & 0 & \hat{\sigma}^2_0 \end{bmatrix} \tag{4.18}$$

(where, once more, $x$ denotes undetermined elements).

Since $\hat{\boldsymbol{\mathcal{H}}}^H \hat{\boldsymbol{R}}_N \hat{\boldsymbol{\mathcal{H}}}$ is a Hermitian matrix, the elements designated by " $x$ " in (4.18) must be equal to zero. Hence, we have proven the following result which is essential in establishing a relation between the AR and Capon methods of spectral estimation:

The parameters $\{\hat{\boldsymbol{a}}_k, \hat{\sigma}^2_k\}$ of the AR models of orders $k = 1, 2, \dots, N-1$ determine the following factorization of the inverse (sample) covariance matrix:

$$\hat{\boldsymbol{R}}_N^{-1} = \hat{\boldsymbol{\mathcal{H}}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mathcal{H}}}^H ; \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{N-1} & & & (0) \\ & \hat{\sigma}^2_{N-2} & & \\ & & \ddots & \\ (0) & & & \hat{\sigma}^2_0 \end{bmatrix} \tag{4.19}$$

Let

$$\hat{A}_k(\nu) = \begin{bmatrix} 1 & e^{-2\imath\pi\nu} & \dots e^{-2\imath\pi k\nu} \end{bmatrix} \begin{bmatrix} 1 \\ -\hat{\boldsymbol{a}}_k \end{bmatrix} \tag{4.20}$$

denote the trigonometric polynomial corresponding to the $k$-th order AR model, and let

$$\hat{S}_{\mathrm{AR}(k),XX}(\nu) = \frac{\hat{\sigma}^2_k}{\left|\hat{A}_k(\nu)\right|^2} \tag{4.21}$$

denote its associated PSD (see Chapter 3). It is readily verified that

$$\begin{aligned} \boldsymbol{e}^H(\nu) \hat{\boldsymbol{\mathcal{H}}} &= \begin{bmatrix} 1 & e^{-2\imath\pi\nu} \dots e^{-2\imath\pi(N-1)\nu} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ & 1 & & \vdots & \vdots \\ & & \ddots & 0 & \vdots \\ & & & 1 & 0 \\ -\hat{\boldsymbol{a}}_{N-1} & -\hat{\boldsymbol{a}}_{N-2} & & -\hat{\boldsymbol{a}}_1 & 1 \end{bmatrix} \\ &= \left[\hat{A}_{N-1}(\nu), e^{-2\imath\pi\nu}\hat{A}_{N-2}(\nu), \dots, e^{-2\imath\pi(N-1)\nu}\hat{A}_0(\nu)\right] \end{aligned} \tag{4.22}$$

It follows from (4.19) and (4.22) that the quadratic form in the denominator of the Capon spectral estimator can be written as

$$\begin{aligned} \boldsymbol{e}^H(\nu) \hat{\boldsymbol{R}}_N^{-1} \boldsymbol{e}(\nu) &= \boldsymbol{e}^H(\nu) \hat{\boldsymbol{\mathcal{H}}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mathcal{H}}}^H \boldsymbol{e}(\nu) \\ &= \sum_{k=0}^{N-1} \frac{\left|\hat{A}_k(\nu)\right|^2}{\hat{\sigma}^2_k} = \sum_{k=0}^{N-1} \frac{1}{\hat{S}_{\mathrm{AR}(k),XX}(\nu)} \end{aligned} \tag{4.23}$$

which finally leads to equation (4.14). □

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# Chapter 5

# Line spectrum estimation

This chapter is devoted to the parametric estimation of a signal composed of a sum of exponentially modulated sinusoids, perturbed by an additive noise. The maximum likelihood principle then reduces the estimation of amplitudes and phases to a simple least squares problem, while the estimation of frequencies and damping factors requires more sophisticated methods, called *high resolution methods*, because they overcome the limits of Fourier analysis in terms of spectral resolution.

The origin of the HR methods dates back to Prony's work published in 1795, which aims to estimate a sum of exponentials by linear prediction techniques [24]. More recently, this approach was further developped by Pisarenko to estimate sinusoids of constant amplitude [25]. In comparison, modern HR methods are based on the particular properties of the signal covariance matrix. Thus, the study of its rank makes it possible to separate the data space into two subspaces, the signal subspace spanned by the sinusoids, and the noise subspace which is its orthogonal complement. The HR methods resulting from this decomposition into subspaces are known to be more robust than linear prediction techniques. This is the case of the *MUltiple SIgnal Classification* (MUSIC) [26] and root-MUSIC [27] methods (which are based on the noise subspace), of the *Toeplitz Approximation Method* (TAM) algorithm [28], as well as the *Estimation of Signal Parameters via Rotational Invariance Techniques* (ESPRIT) algorithm [29] and its variants TLS-ESPRIT [30] and PRO-ESPRIT [31] (which are based on the signal subspace). All of these estimation methods can be applied to the ESM (Exponential Sinusoidal Model), which represents the signal as a sum of exponentially modulated sinusoids. This model is also called *Exponentially Damped Sinusoids* (EDS) when the modulation is decreasing [32]. Other estimation techniques have been specifically developed for the ESM model, such as the *Kumaresan and Tufts* (KT) algorithm, also called Min-Norm method [33], its modified version [34] (based on linear prediction), and the *Matrix Pencil* method [35] (based on subspaces). A more complete list of methods can be found in [36].

This chapter is not intended to present the HR methods exhaustively, but rather to familiarize the reader with the concepts on which they are based. This is why only some of them are presented here: the Prony, Pisarenko, MUSIC and ESPRIT methods. This presentation will start with the definition of the signal model (Section 1). Then the maximum likelihood method, which makes it possible to establish a link with the Fourier transform, will be presented in Section 2. Then the high resolution methods that estimate the complex poles will be introduced in Section 3, and techniques for estimating the other parameters of the model will be presented in Section 4. Section 5 will be devoted to the analysis of the performance of HR methods. Finally, the results of this chapter will be summarized in Section 6.

## 1 Signal model

Consider the discrete signal model (defined for all $t \in \mathbb{Z}$)

$$\boxed{s(t) = \sum_{k=0}^{K-1} \alpha_k z_k{}^t} \tag{5.1}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

where $K \in \mathbb{N}^*$, $\forall k \in \{0 \dots K-1\}$, $\alpha_k \in \mathbb{C}^*$, and the poles $z_k \in \mathbb{C}^*$ are pairwise distinct. In the particular case where all the poles belong to the unit circle, the signal is represented as a sum of complex sinusoids. Thus, each pole $z_k$ is written in the form $z_k = e^{2\imath\pi f_k}$ where $f_k \in \mathbb{R}$ is the frequency of the sinusoid. More generally, if the poles are not on the unit circle, the sinusoids are exponentially modulated (ESM). In this case, each pole $z_k$ is written in polar form $z_k = e^{\delta_k}\, e^{i2\pi f_k}$, where $\delta_k \in \mathbb{R}$ is the damping factor (or attenuation rate) of the sinusoid. In particular, poles with the same polar angle and different modules are associated with the same frequency. The complex amplitudes $\alpha_k$ are also written in polar form $\alpha_k = a_k\, e^{\imath\phi_k}$, where $a_k \in \mathbb{R}_+^*$ and $\phi \in \mathbb{R}$.

In addition, the observed signal $x(t)$ can be modeled as the sum of the deterministic signal $s(t)$ defined above and of a complex centered white Gaussian noise $b(t)$ of variance $\sigma^2$. Remember that a complex centered white Gaussian noise is a sequence of *i.i.d* random variables with complex values, of probability density $p(b) = \frac{1}{\pi\sigma^2}\, e^{-\frac{|b|^2}{\sigma^2}}$. We thus obtain the relation

$$x(t) = s(t) + b(t). \tag{5.2}$$

The signal is observed over time windows of length $N \geq K$. Thus, for all $t \in \mathbb{Z}$, we consider the time window $\{t-l+1 \dots t+n-1\}$, where the integers $n$ and $l$ are such that $N = n + l - 1$, and we define the vector $\boldsymbol{s}(t) = [s(t-l+1), \dots, s(t+n-1)]^\top$, of dimension $N$. For all $z \in \mathbb{C}$, let us define $\boldsymbol{v}(z) = [1, z, \dots, z^{N-1}]^\top$. However $\boldsymbol{s}(t) = \sum_{k=0}^{K-1} \alpha_k\, z_k^{t-l+1}\, \boldsymbol{v}(z_k)$ . This equality can be rewritten in the form of a product: $\boldsymbol{s}(t) = \boldsymbol{V}^N\, \boldsymbol{D}^{t-l+1}\, \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\alpha_0, \dots, \alpha_{K-1}]^\top$ is a vector of dimension $K$, $\boldsymbol{D} = \mathrm{diag}(z_0, \dots, z_{(K-1)})$ is a diagonal matrix of dimension $K \times K$, and $\boldsymbol{V}^N = [\boldsymbol{v}(z_0), \dots, \boldsymbol{v}(z_{(K-1)})]$ is a Vandermonde matrix of dimensions $N \times K$: (*cf.* definition 19 in appendix 7.2 page 56)

$$\boldsymbol{V}^N = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ \vdots & \vdots & \vdots & \vdots \\ {z_0}^{N-1} & {z_1}^{N-1} & \dots & {z_{K-1}}^{N-1} \end{bmatrix}.$$

Then define the vector of amplitudes at time $t$, $\boldsymbol{\alpha}(t) = \boldsymbol{D}^{t-l+1}\boldsymbol{\alpha}$, so that $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$. It is known that the determinant of the square Vandermonde matrix $\boldsymbol{V}^K$ extracted from the first $K$ rows of $\boldsymbol{V}^N$ (remember that $N \geq K$) is (*cf.* proposition 38 in appendix 7.2 page 56)

$$\boxed{\det(\boldsymbol{V}^K) = \prod_{0 \leq k_1 < k_2 \leq K-1} (z_{k_2} - z_{k_1}).} \tag{5.3}$$

Thus, matrix $\boldsymbol{V}^N$ has full rank if and only if all the poles are distinct. The relation $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$ therefore shows that for each time $t$ the vector $\boldsymbol{s}(t)$ lies in the range space of matrix $\boldsymbol{V}^N$, of dimension less than or equal to $K$ in the general case, and equal to $K$ if all poles are distinct.

Let $\boldsymbol{b}(t) = [b(t-l+1), \dots, b(t+n-1)]^\top$ be the vector containing the samples of the additive noise. It is a centered Gaussian random vector, whose covariance matrix is $\boldsymbol{R}_{bb} = \sigma^2\boldsymbol{I}_N$. Finally, let $\boldsymbol{x}(t) = [x(t-l+1), \dots, x(t+n-1)]^\top$ be the vector of observed data. This vector therefore satisfies $\boldsymbol{x}(t) = \boldsymbol{s}(t) + \boldsymbol{b}(t)$. The model being posed, the analysis of the signal $s(t)$ will consist in estimating the parameters $\sigma^2$, $z_0$, $\dots$, $z_{(K-1)}$ and $\boldsymbol{\alpha}(t)$. A classical parametric estimation technique, the maximum likelihood method, is applied to this model in the next section.

# 2 Maximum likelihood method

The maximum likelihood principle is a general parametric estimation method. It provides asymptotically efficient and unbiased estimators. This is why it is often preferred to other estimation techniques when it has a simple closed-form solution.

## 2.1 Application of the maximum likelihood principle to the ESM model

The maximum likelihood principle consists in maximizing the conditional probability of observing the signal $x$ over the interval $\{t-l+1, \dots, t+n-1\}$, given the parameters $\sigma^2$, $z_0$, $\dots$, $z_{(K-1)}$ and $\boldsymbol{\alpha}(t)$ (or the natural logarithm

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

of this probability, called *log-likelihood* of the observations). Since $\boldsymbol{x}(t) = \boldsymbol{s}(t) + \boldsymbol{b}(t)$, where $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$ is a deterministic vector and $\boldsymbol{b}(t)$ is a centered complex Gaussian random vector of covariance matrix $\boldsymbol{R}_{bb} = \sigma^2\boldsymbol{I}_N$, $\boldsymbol{x}(t)$ is itself a complex Gaussian random vector with expected value $\boldsymbol{s}(t)$ and covariance matrix $\boldsymbol{R}_{bb}$. Remember that the probability density of such a random vector is

$$p(\boldsymbol{x}(t)) = \frac{1}{\pi^N \det(\boldsymbol{R}_{bb})}\, e^{-(\boldsymbol{x}(t)-\boldsymbol{s}(t))^H \boldsymbol{R}_{bb}^{-1}(\boldsymbol{x}(t)-\boldsymbol{s}(t))}.$$

So the log-likelihood of the observations is

$$L(\sigma^2, z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) = -N\ln(\pi\sigma^2) - \frac{1}{\sigma^2}\, g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$$

where

$$g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) = \left(\boldsymbol{x}(t) - \boldsymbol{V}^N\boldsymbol{\alpha}(t)\right)^H \left(\boldsymbol{x}(t) - \boldsymbol{V}^N\boldsymbol{\alpha}(t)\right).$$

Maximizing this log-likelihood with respect to the parameters $(\sigma^2, z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$ can be done by first minimizing $g$ with respect to the pair $(z_0 \dots z_{K-1},\ \boldsymbol{\alpha}(t))$, then by maximizing $L$ with respect to $\sigma$. We thus obtain $\sigma^2 = \frac{1}{N}\, g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$, i.e.

$$\boxed{\sigma^2 = \tfrac{1}{N}\, \left\|\boldsymbol{x}(t) - \boldsymbol{V}^N\boldsymbol{\alpha}(t)\right\|^2.} \tag{5.4}$$

It appears that $\sigma^2$ is estimated by calculating the power of the residual obtained by subtracting the exponentials from the observed signal.

The matrix $\boldsymbol{V}^N$ has full rank, since it has been assumed in section 1 that the poles are pairwise distinct. Thus, matrix $\boldsymbol{V}^{N\,H}\boldsymbol{V}^N$ is invertible. In order to minimize $g$ with respect to the pair $(z_0 \dots z_{K-1},\ \boldsymbol{\alpha}(t))$, we just use the decomposition

$$\begin{aligned} g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) &= \boldsymbol{x}(t)^H\boldsymbol{x}(t) - \boldsymbol{x}(t)^H\boldsymbol{V}^N \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right)^{-1} \boldsymbol{V}^{N\,H}\boldsymbol{x}(t) \\ &+ \left(\boldsymbol{\alpha}(t) - \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right)^{-1} \boldsymbol{V}^{N\,H}\boldsymbol{x}(t)\right)^H \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right) \left(\boldsymbol{\alpha}(t) - \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right)^{-1} \boldsymbol{V}^{N\,H}\boldsymbol{x}(t)\right). \end{aligned}$$

The last term of this equation is always non-negative, and can be zeroed by defining

$$\boxed{\boldsymbol{\alpha}(t) = \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right)^{-1} \boldsymbol{V}^{N\,H}\boldsymbol{x}(t).} \tag{5.5}$$

It appears that the vector of complex amplitudes $\boldsymbol{\alpha}(t)$ is estimated in the same way as with the ordinary least squares method.

Function $g$ is therefore minimal when the $K$ -tuple $(z_0 \dots z_{K-1})$ maximizes function $\mathcal{J}$ defined by

$$\boxed{\mathcal{J}(z_0, \dots, z_{(K-1)}) = \boldsymbol{x}(t)^H\boldsymbol{V}^N \left(\boldsymbol{V}^{N\,H}\boldsymbol{V}^N\right)^{-1} \boldsymbol{V}^{N\,H}\boldsymbol{x}(t).} \tag{5.6}$$

As this optimization problem does not have a closed-form solution in the general case, it must be solved numerically. In summary, the maximum likelihood principle leads to estimating the parameters of the model in three stages:

**complex poles** are obtained by maximizing function $\mathcal{J}$ (equation (5.6)),

**complex amplitudes** are obtained by calculating the right side of equation (5.5),

**the standard deviation** is then given by equation (5.4).

Unfortunately, it turns out that the first step of this estimation method, which requires the optimization of a function of $K$ complex variables, is difficult to implement, because the function to be maximized has many local maxima. In addition, it is extremely costly in terms of computation time. This is why we generally use more reliable and faster methods to estimate complex poles. However, once the poles are estimated, the maximum likelihood principle can be used to determine the complex amplitudes and the standard deviation of the noise.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

## 2.2 Maximum likelihood and Fourier resolution

Let us now take a look at the particular case where all the poles are on the unit circle ($\forall k,\ \delta_k = 0$). The results of section 2.1 showed that the maximum likelihood principle leads to an optimization problem which does not have a simple closed-form solution in the general case. However, such a solution exists in the particular case where $K = 1$, as well as an approximate solution if $K > 1$.

Let us first examine the case of a single complex exponential ($K = 1$). Then equation (5.6) is simplified as $\mathcal{J}(z_0) = \widehat{R}_x(z_0)$, where $\widehat{R}_x$ is the periodogram of the signal $x(t)$ observed on the time window $\{t-l+1 \dots t+n-1\}$:

$$\widehat{R}_x(e^{i2\pi f_0}) = \frac{1}{N}\left|X(e^{i2\pi f_0})\right|^2$$

where $X(e^{i2\pi f_0}) = \boldsymbol{v}(e^{i2\pi f_0})^H \boldsymbol{x}(t) = \sum_{\tau=0}^{N-1} x(t-l+1+\tau)\, e^{-i2\pi f_0 \tau}$. Similarly, equation (5.5) is simplified as $\alpha_0(t) = \frac{1}{N} X(e^{i2\pi f_0})$. Finally, equation (5.4) is simplified as $\sigma^2 = \frac{1}{N}\left(\|\boldsymbol{x}(t)\|^2 - \widehat{R}_x(e^{i2\pi f_0})\right)$.

These results lead to the following conclusion:

> The maximum likelihood principle leads in the case of a complex sinusoid to detect the frequency at which the periodogram reaches its maximum. The corresponding complex amplitude is proportional to the value of the DFT of the signal at this frequency. The noise variance is estimated as the signal average power after subtracting the sinusoid.

Let us now address the general case $K \geq 1$, for which the maximization of function $\mathcal{J}(z)$ no longer has an exact closed-form solution. We then introduce the following hypothesis:

$$N >> \frac{1}{\min\limits_{k_1 \neq k_2} \left|f_{k_2} - f_{k_1}\right|}.$$

Matrix $\boldsymbol{V^N}^H \boldsymbol{V^N}$ is a positive definite Hermitian matrix of dimension $K \times K$, whose entries can be calculated in closed-form: $\left\{\boldsymbol{V^N}^H \boldsymbol{V^N}\right\}_{(k_1,k_2)} = \sum_{\tau=0}^{N-1} (\overline{z_{k_1}} z_{k_2})^\tau$. We then obtain

$$\begin{array}{lcll} \frac{1}{N}\left\{\boldsymbol{V^N}^H \boldsymbol{V^N}\right\}_{(k_1,k_2)} & = & e^{\imath\pi(N-1)(f_{k_2}-f_{k_1})} \frac{\sin(\pi N(f_{k_2}-f_{k_1}))}{N\sin(\pi(f_{k_2}-f_{k_1}))} & \text{if } k_1 \neq k_2 \\ \frac{1}{N}\left\{\boldsymbol{V^N}^H \boldsymbol{V^N}\right\}_{(k,k)} & = & 1 & \text{if } k_1 = k_2 = k \end{array}$$

Therefore, when $N >> \frac{1}{\min\limits_{k_1 \neq k_2} \left|f_{k_2} - f_{k_1}\right|}$, $\frac{1}{N}\boldsymbol{V^N}^H \boldsymbol{V^N} = \boldsymbol{I}_K + O\left(\frac{1}{N}\right)$, thus

$$\left(\boldsymbol{V^N}^H \boldsymbol{V^N}\right)^{-1} = \frac{1}{N}\boldsymbol{I}_K + O\left(\frac{1}{N^2}\right).$$

Then equation (5.6) is simplified as

$$\mathcal{J}(z_0, \dots, z_{K-1}) = \frac{1}{N}\left\|\boldsymbol{V^N}^H \boldsymbol{x}(t)\right\|^2 + O\left(\frac{1}{N^2}\right) = \sum_{k=0}^{K-1} \widehat{R}(z_k) + O\left(\frac{1}{N^2}\right).$$

Similarly, equation (5.5) is simplified as $\boldsymbol{\alpha}(t) = \frac{1}{N}\boldsymbol{V^N}^H \boldsymbol{x}(t) + O\left(\frac{1}{N^2}\right)$, hence

$$\alpha_k(t) = \frac{1}{N} X(e^{i2\pi f_k}) + O\left(\frac{1}{N^2}\right).$$

Finally, equation (5.4) is simplified as $\sigma^2 = \frac{1}{N}\left(\|\boldsymbol{x}(t)\|^2 - \sum_{k=0}^{K-1} \widehat{R}(e^{i2\pi f_k})\right) + O\left(\frac{1}{N^2}\right)$.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Thus, the joint maximization of $\mathcal{J}$ with respect to $z_0, \ldots, z_{K-1}$ leads to determine the $K$ frequencies corresponding to the $K$ largest values of the periodogram. The corresponding complex amplitudes are proportional to the value of the DFT of the signal at these frequencies. Remember that these results are only valid if all the poles are on the unit circle and are based on the assumption $N >> \frac{1}{\min\limits_{k_1 \neq k_2} |f_{k_2} - f_{k_1}|}$.

We thus observe the limit of the Fourier analysis in terms of spectral resolution: the parameters are estimated correctly provided that the length of the observation window is sufficiently large compared to the inverse of the smallest frequency difference between two neighboring poles. It is this limit that the HR methods presented in section 3 allow to overcome. So, HR methods are able to distinguish two close sinusoids, that Fourier analysis does not allow to distinguish. In applications, HR methods can be used with shorter windows than those usually used with Fourier analysis.

Figure 5.1: Jean Baptiste Joseph FOURIER (1768-1830)

# 3 High resolution methods

We begin here by introducing the oldest high-resolution methods, which are based on linear prediction techniques (section 3.1), before addressing in section 3.2 the more recent subspace methods.

## 3.1 Linear prediction techniques

The first two high-resolution methods presented in this chapter are based on a fundamental result related to linear recurrence equations, presented in section 3.1.1.

### 3.1.1 Linear recurrence equations

Let $p_0 \in \mathbb{C}^*$, $K \in \mathbb{N}^*$, and $\{z_0, \ldots, z_{K-1}\}$ be $K$ distinct and non-zero complex numbers. We define the polynomial of degree $K$ whose dominant coefficient is $p_0$ and whose roots are $z_k$:

$$P[z] = p_0 \prod_{k=0}^{K-1} (z - z_k) = \sum_{\tau=0}^{K} p_{K-\tau}\, z^{\tau}.$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

The following theorem characterizes the signal model.

**Theorem 30.** *A complex discrete signal $\{s(t)\}_{t\in\mathbb{Z}}$ satisfies the recurrence equation*

$$\boxed{\sum_{\tau=0}^{K} p_\tau\, s(t-\tau) = 0} \tag{5.7}$$

*for all $t \in \mathbb{Z}$ if and only if there are scalars $\alpha_0, \ldots, \alpha_{K-1} \in \mathbb{C}$ such that $s(t) = \sum_{k=0}^{K-1} \alpha_k\, {z_k}^t$.*

*Proof.* First of all, it is straightforward to check that the set of signals which satisfy the relation (5.7) forms a vector space $E$ over $\mathbb{C}$. Next, we will prove that this vector space has dimension less than or equal to $K$. Consider the application

$$\begin{array}{rccc} f: & E & \to & \mathbb{C}^K \\ & s[t] & \mapsto & [s[0], \ldots, s[K-1]]^\top \end{array}$$

We can notice that $f$ is a linear map. Let $s \in E$ be a signal such that $f(s) = \mathbf{0}$. Then $s$ is zero over the interval $[0, K-1]$. By using the recurrence (5.7), we deduce that $s$ is also zero over the interval $[K, +\infty[$. Finally, using the recurrence (5.7) and the fact that $p_K \neq 0$, we show that $s$ is also zero over the interval $]-\infty, -1]$. Consequently, $s \equiv 0$, so the linear map $f$ is injective. We conclude that the vector space $E$ is at most of dimension $K$.

Now we will show that any signal of the form $s[t] = {z_k}^t$ where $k \in \{0, \ldots, K-1\}$ belongs to the vector space $E$. Indeed, if $s[t] = {z_k}^t$, then $\forall t \in \mathbb{Z}$, $\sum_{k=0}^{K} p_k\, s[t-k] = {z_k}^{t-K} \sum_{k=0}^{K} p_k\, {z_k}^{K-k} = {z_k}^{t-K}\, P[z_k] = 0$, therefore $s[t]$ satisfies the relationship (5.7).

Finally, consider the family of vectors $\{{z_k}^t\}_{\{k\in\{0,\ldots,K-1\}\}}$. The square matrix whose columns are extracted from these vectors and whose rows correspond to times $\{0 \ldots K-1\}$ is a Vandermonde matrix (*cf.* definition 19 of the appendix 7.2 page 56). According to proposition 38 in appendix 7.2, it is invertible, since the poles $z_k$ are pairwise distinct. Therefore, the family $\{{z_k}^t\}_{\{k\in\{0,\ldots,K-1\}\}}$ is linearly independent. However it contains precisely $K$ vectors of $E$. This vector space is therefore exactly of dimension $K$, and this family forms a basis of it. Thus, a signal $s[t]$ belongs to $E$ if and only if it is of the form (5.2). □ □

### 3.1.2 Prony method

The work of Baron de Prony is at the origin of the development of high resolution methods. He proposed an estimation method inspired by the previous result on the linear recurrence equations [24]. This method was originally intended to estimate noiseless real exponentials; however we apply it here to the estimation of noisy complex exponentials. Prony's method consists in first determining the polynomial $P[z]$ using linear prediction techniques, then extracting the roots of this polynomial. We define the prediction error

$$\varepsilon(t) \triangleq \sum_{\tau=0}^{K} p_\tau\, x(t-\tau). \tag{5.8}$$

In particular, by substituting equations (5.2) and (5.7) into equation (5.8), we get $\varepsilon(t) = \sum_{\tau=0}^{K} p_\tau\, b(t-\tau)$. The prediction error therefore characterizes only the noise which is superimposed on the signal. Let us address the particular case $n = K+1$, and suppose that $l \geq K+1$. Thus, the signal is observed on the window $\{t-l+1 \ldots t+K\}$. By applying equation (5.8) at times $\{t-l+K+1,\ t-l+K+2,\ \ldots,\ t+K\}$, we get the system of equations

$$\left\{\begin{array}{ccccccccc} p_0\, x(t-l+K+1) & + & p_1\, x(t-l+K) & + & \ldots & + & p_K\, x(t-l+1) & = & \varepsilon(t-l+K+1) \\ p_0\, x(t-l+K+2) & + & p_1\, x(t-l+K+1) & + & \ldots & + & p_K\, x(t-l+2) & = & \varepsilon(t-l+K+2) \\ \vdots & + & \vdots & + & \ldots & + & \vdots & = & \vdots \\ p_0\, x(t+K) & + & p_1\, x(t+K-1) & + & \ldots & + & p_K\, x(t) & = & \varepsilon(t+K) \end{array}\right. \tag{5.9}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Figure 5.2: Gaspard-Marie RICHE de PRONY (1755-1839)

Then define $\boldsymbol{p} = [p_K,\ p_{(K-1)}, \ldots, p_0]^H$, $\boldsymbol{\varepsilon}(t) = [\varepsilon(t-l+K+1), \varepsilon(t-l+K+2), \ldots, \varepsilon(t+K)]^H$ and

$$\boldsymbol{X}(t) = \begin{bmatrix} x(t-l+1) & \cdots & x(t-1) & x(t) \\ x(t-l+2) & \cdots & x(t) & x(t+1) \\ \vdots & \cdots & \vdots & \vdots \\ x(t-l+K+1) & \cdots & x(t+K-1) & x(t+K) \end{bmatrix} \tag{5.10}$$

so that the system of equations (5.9) can be condensed in the form $\boldsymbol{p}^H \boldsymbol{X}(t) = \boldsymbol{\varepsilon}(t)^H$.

Prony's method consists in minimizing the power of the prediction error $\frac{1}{l}\,\|\boldsymbol{\varepsilon}\|^2$ with respect to $\boldsymbol{p}$, subject to the constraint $p_0 = 1$. However it is possible to write $\frac{1}{l}\|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{p}^H \widehat{\boldsymbol{R}}_{xx}(t)\,\boldsymbol{p}$, where matrix $\widehat{\boldsymbol{R}}_{xx}(t) = \frac{1}{l}\,\boldsymbol{X}(t)\,\boldsymbol{X}(t)^H$ has dimension $(K+1)\times(K+1)$. As matrix $\boldsymbol{X}(t)$ has $K+1$ rows and $l \geq K+1$ columns, we can assume that matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ is invertible.

Theorem 37 in appendix 7.1 page 56 allows to prove [1] that the solution of this optimization problem is [2]

$$\boldsymbol{p} = \frac{1}{\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1}\, \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$$

where $\boldsymbol{e}_1 \triangleq [1,\ 0 \ldots 0]^\top$ is a vector of dimension $K+1$. Thus, the Prony estimation method includes the following steps:

- Construct matrix $\boldsymbol{X}(t)$ and calculate $\widehat{\boldsymbol{R}}_{xx}(t)$;
- Compute $\boldsymbol{p} = \frac{1}{\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1}\, \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$;
- Determine the poles $\{z_0, \ldots, z_{K-1}\}$ as the roots of the polynomial $P[z] = \sum_{k=0}^{K} p_k\, z^{K-k}$.

[1] As the data are complex, it is necessary to decompose the vector $\boldsymbol{p}$ into its real part and its imaginary part to be able to apply theorem 37, which deals exclusively with the real data case.

[2] The scalar $\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$ is non-zero, since the vector $\boldsymbol{e}_1$ is unitary and matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ is positive definite.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

### 3.1.3 Pisarenko method

The Pisarenko method is a variant of the Prony method. It consists in minimizing the power of the prediction error $\frac{1}{l}\|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{p}^H \widehat{\boldsymbol{R}}_{xx}(t)\,\boldsymbol{p}$ subject to the constraint that vector $\boldsymbol{p}$ has norm 1. Theorem 37 in appendix 7.1 page 56 allows us to prove that the solution of this optimization problem is the eigenvector of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ associated with the smallest eigenvalue.

Thus the Pisarenko method [25] consists of the following stages:

- calculate and diagonalize $\widehat{\boldsymbol{R}}_{xx}(t)$;
- determine $\boldsymbol{p}$ as the eigenvector associated with the smallest eigenvalue;
- extract the roots of polynomial $P[z]$.

The Prony and Pisarenko methods are the oldest HR methods. As we will show in section 5.2, they do not prove to be very robust in practice, this is why the subspace methods, proposed more recently, are generally preferred to them.

## 3.2 Subspace methods

In the same spirit as the Pisarenko method, modern HR methods (for example [26, 29, 35]) are based on a decomposition of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$.

### 3.2.1 Singular structure of the data matrix

Suppose now that $n \geq K+1$ and $l \geq K+1$, and construct the data matrix of the noiseless signal $s(t)$ on the same model as matrix $\boldsymbol{X}(t)$ in equation (5.10), according to a Hankel structure:

$$\boldsymbol{S}(t) = \begin{bmatrix} s(t-l+1) & \cdots & s(t-1) & s(t) \\ s(t-l+2) & \cdots & s(t) & s(t+1) \\ \vdots & \cdots & \vdots & \vdots \\ s(t-l+n) & \cdots & s(t+n-2) & s(t+n-1) \end{bmatrix}. \tag{5.11}$$

The following proposition characterizes the signal model.

**Proposition 31** (Factorization of the data matrix)**.** *The following assertions are equivalent:*

1. *The signal $s(t)$ satisfies the model defined in equation* (5.1) *on the interval $\{t-l+1, \ldots, t+n-1\}$;*
2. *The matrix $\boldsymbol{S}(t)$ defined in equation* (5.11) *can be factorized as*

$$\boxed{\boldsymbol{S}(t) = \boldsymbol{V}^n\,\boldsymbol{A}(t)\,\boldsymbol{V}^{l\top}} \tag{5.12}$$

*where the diagonal matrix $\boldsymbol{A}(t) = \operatorname{diag}(z_0^{t-l+1}\alpha_0, \ldots, z_{(K-1)}^{t-l+1}\alpha_{(K-1)})$ has dimension $K \times K$, $\boldsymbol{V}^n$ has dimension $n \times K$, and $\boldsymbol{V}^l$ has dimensions $l \times K$.*

*Proof of Proposition 31:* Let us prove each of the two implications.

**Proof of 1. $\Rightarrow$ 2.** The signal $s(t)$ is defined as a sum of complex exponentials: $s(t) = \sum_{k=0}^{K-1} s_k(t)$, where $s_k(t) = \alpha_k\, z_k{}^t$. Consequently, matrix $\boldsymbol{S}(t)$ defined in equation (5.11) can be decomposed in the same way: $\boldsymbol{S}(t) = \sum_{k=0}^{K-1} \boldsymbol{S}_k(t)$, where matrices $\boldsymbol{S}_k(t)$ are constructed in the same way as $\boldsymbol{S}(t)$ in equation (5.11), from the signals $s_k(t)$. However it is straightforward to check that $\boldsymbol{S}_k(t) = \alpha_k\, z_k^{t-l+1}\, \boldsymbol{v}^n(z_k)\, \boldsymbol{v}^l(z_k)^\top$, where $\boldsymbol{v}^n(z) = [1, z, \ldots, z^{n-1}]^\top$ and $\boldsymbol{v}^l(z) = [1, z, \ldots, z^{l-1}]^\top$. Equation (5.12) follows directly.

**Proof of 1. $\Rightarrow$ 2.** If matrix $\boldsymbol{S}(t)$ is defined by equation (5.12), the reverse reasoning allows to show that $\boldsymbol{S}(t)$ can be written in the form (5.11), where $s(t)$ is defined in equation (5.1). □ □

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Proposition 31 induces a result on which all subspace methods are based:

**Corollary 32.** *Matrix $\boldsymbol{S}(t)$ defined in equation* (5.11) *has rank less than or equal to $K$. More precisely, it has rank $K$ if and only if $n \geq K$, $l \geq K$, all the poles $z_k$ are distinct and non-zero, and all the amplitudes $\alpha_k$ are non-zero. In this case, its range space is spanned by matrix $\boldsymbol{V}^n$.*

*Proof of Corollary 32:* It turns out that matrices $\boldsymbol{V}^n$ and $\boldsymbol{V}^l$ have full rank equal to $K$. Indeed, if $\boldsymbol{V}_0$ is the Vandermonde matrix made up of the first $K$ rows of $\boldsymbol{V}^n$ or $\boldsymbol{V}^l$, proposition 38 shows that $\boldsymbol{V}_0$ is invertible, since the poles $z_k$ are pairwise distinct. Consequently, $\text{rank}(\boldsymbol{V}^n) \geq K$ and $\text{rank}(\boldsymbol{V}^l) \geq K$. However $\dim(\boldsymbol{V}^n) = n \times K$ and $\dim(\boldsymbol{V}^l) = l \times K$, therefore $\text{rank}(\boldsymbol{V}^n) = \text{rank}(\boldsymbol{V}^l) = K$. Furthermore, matrix $\boldsymbol{A}(t)$, of dimensions $K \times K$, is invertible, hence of rank $K$.

We can deduce from this remark that matrix $\boldsymbol{S}(t)$ is also of rank $K$. To do this, let us first show that $\ker(\boldsymbol{S}(t)) = \ker({\boldsymbol{V}^l}^\top)$. Indeed,

- $\forall \boldsymbol{y} \in \mathbb{C}^n$, ${\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0} \Rightarrow \boldsymbol{V}^n \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$;
- $\forall \boldsymbol{y} \in \mathbb{C}^n$, $\boldsymbol{V}^n \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0} \Rightarrow \left({\boldsymbol{V}^n}^H \boldsymbol{V}^n\right) \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$. However matrix $\left({\boldsymbol{V}^n}^H \boldsymbol{V}^n\right) \boldsymbol{A}(t)$ is invertible, hence ${\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$.

The rank-nullity theorem then implies: $\text{rank}(\boldsymbol{S}(t)) = n - \dim(\ker(\boldsymbol{S}(t))) = n - \dim(\ker({\boldsymbol{V}^l}^\top)) = \text{rank}({\boldsymbol{V}^l}^\top) = K$. □ □

The singular structure of the data matrix induces an equivalent structure for the correlation matrix, defined below.

### 3.2.2 Singular structure of the correlation matrix

The subspace methods are based on the particular structure of the signal correlation matrix $\boldsymbol{C}_{ss}(t) = \boldsymbol{S}(t)\,\boldsymbol{S}(t)^H$, and in particular on its eigensubspaces, that we will now study. Let us define $\boldsymbol{R}_{ss}(t) = \frac{1}{l}\,\boldsymbol{C}_{ss}(t)$. Equation (5.12) shows that

$$\boxed{\boldsymbol{R}_{ss}(t) = \boldsymbol{V}^n\,\boldsymbol{P}(t)\,{\boldsymbol{V}^n}^H} \tag{5.13}$$

where

$$\boldsymbol{P}(t) = \frac{1}{l}\,\boldsymbol{A}(t)\,{\boldsymbol{V}^l}^\top \overline{\boldsymbol{V}^l}\,\boldsymbol{A}(t)^H \tag{5.14}$$

is a symmetric positive definite matrix. Thus, equation (5.13) shows that under the same assumptions as for $\boldsymbol{S}(t)$, matrix $\boldsymbol{R}_{ss}(t)$ has rank $K$.

The range space of matrix $\boldsymbol{R}_{ss}(t)$, of dimension $K$, is spanned by matrix $\boldsymbol{V}^n$. This vector space is called *signal subspace* in the literature.

Then let $\{\boldsymbol{w}_m\}_{m=0\ldots n-1}$ be an orthonormal basis of eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$, associated to the eigenvalues $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1} \geq 0$. Since matrix $\boldsymbol{R}_{ss}(t)$ has only of rank $K$, we actually have $\lambda_m = 0\ \forall m \geq K$. We denote by $\boldsymbol{W}(t)$ matrix $[\boldsymbol{w}_0 \ldots \boldsymbol{w}_{K-1}]$, and by $\boldsymbol{W}_\perp(t)$ matrix $[\boldsymbol{w}_K \ldots \boldsymbol{w}_{n-1}]$. We can then check that ${}^\top\mathcal{I}\text{m}(\boldsymbol{W}(t)) = {}^\top\mathcal{I}\text{m}(\boldsymbol{V}^n)$. Indeed, vectors $\{\boldsymbol{w}_m\}_{m=0\ldots K-1}$ are eigenvectors of matrix $\boldsymbol{R}_{ss}(t) = \boldsymbol{V}^n\boldsymbol{P}(t){\boldsymbol{V}^n}^H$ associated with nonzero eigenvalues. So $\forall\, k \in \{0 \ldots K-1\}$, $\boldsymbol{w}_m \in {}^\top\mathcal{I}\text{m}(\boldsymbol{V}^n)$. Thus ${}^\top\mathcal{I}\text{m}(\boldsymbol{W}(t)) \subset {}^\top\mathcal{I}\text{m}(\boldsymbol{V}^n)$. However matrices $\boldsymbol{W}(t)$ and $\boldsymbol{V}^n$ have the same rank $K$, therefore ${}^\top\mathcal{I}\text{m}(\boldsymbol{W}(t)) = {}^\top\mathcal{I}\text{m}(\boldsymbol{V}^n)$.

Matrix $\boldsymbol{W}(t)$ is another basis of the signal subspace, generally distinct from $\boldsymbol{V}^n$.

We then define matrix $\boldsymbol{X}(t)$ from the samples of the noisy signal $x(t)$, in the same way as matrix $\boldsymbol{S}(t)$ in equation (5.11), and we consider the correlation matrix

$$\boldsymbol{C}_{xx}(t) = \boldsymbol{X}(t)\,\boldsymbol{X}(t)^H. \tag{5.15}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Then let $\widehat{\boldsymbol{R}}_{xx}(t) = \frac{1}{l}\boldsymbol{C}_{xx}(t)$ (as in section 3.1.2). Since the additive noise $b(t)$ is white and centered, of variance $\sigma^2$, matrix $\boldsymbol{R}_{xx}(t) = \mathbb{E}[\widehat{\boldsymbol{R}}_{xx}(t)]$ is such that

$$\boldsymbol{R}_{xx}(t) = \boldsymbol{R}_{ss}(t) + \sigma^2 \boldsymbol{I}_n. \tag{5.16}$$

Using equation (5.16), we show that the family $\{\boldsymbol{w}_m\}_{m=0\ldots n-1}$ defined above is also an orthonormal basis of eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$, associated with the eigenvalues

$$\overline{\lambda}_m = \begin{cases} \lambda_m + \sigma^2 \ \forall m \in \{0, \ldots K-1\} \\ \sigma^2 \ \forall m \in \{K, \ldots n-1\} \end{cases}.$$

Thus, all the eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$ are also eigenvectors of $\boldsymbol{R}_{xx}(t)$, and the corresponding eigenvalues of $\boldsymbol{R}_{xx}(t)$ are equal to those of $\boldsymbol{R}_{ss}(t)$ plus $\sigma^2$. Consequently, the signal subspace, defined as the range space of matrix $\boldsymbol{R}_{ss}(t)$, is also the principal subspace of dimension $K$ of matrix $\boldsymbol{R}_{xx}(t)$, i.e. the eigensubspace of $\boldsymbol{R}_{xx}(t)$ associated with the $K$ largest eigenvalues, all strictly greater than $\sigma^2$. The $n-K$ eigenvalues associated with the orthogonal complement of the signal subspace, called *noise subspace*, are all equal to $\sigma^2$. Thus, it is possible to estimate the signal subspace and the noise subspace by calculating the *EigenValue Decomposition* (EVD) of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$, or even the *Singular Value Decomposition* (SVD) of $\boldsymbol{X}(t)$. By concatenating the $K$ principal eigen or singular vectors of one of these matrices, we thus obtain matrix $\boldsymbol{W}(t) = [\boldsymbol{w}_0 \ldots \boldsymbol{w}_{K-1}]$ of dimensions $n \times K$ spanning the signal subspace, and by concatenating the $n-K$ other vectors, we obtain matrix $\boldsymbol{W}_\perp(t) = [\boldsymbol{w}_K \ldots \boldsymbol{w}_{n-1}]$ of dimensions $n \times (nK)$ spanning the noise subspace.

The idea of decomposing the data space into two subspaces (signal and noise) is the source of several high-resolution methods, including the MUSIC method, presented in section 3.2.4, and the ESPRIT method, presented in section 3.2.5.

### 3.2.3 Complement: analogy between the spectrum in the matrix sense and in the Fourier sense

We examine here the particular case where all poles are on the unit circle ($\forall k,\ \delta_k = 0$) and where all frequencies $f_k$ are both multiple of $\frac{1}{n}$ and $\frac{1}{l}$ (we will consider the results that we will get as asymptotic results). We denote by $X\left(e^{i2\pi\frac{\nu}{n}}\right)$ the DFT of the signal observed on the window $\{t, \ldots, t+n-1\}$:

$$X\left(e^{i2\pi\frac{\nu}{n}}\right) = \sum_{\tau=0}^{n-1} x(t+\tau)\, e^{-i2\pi\frac{\nu}{n}}.$$

We then define the periodogram of the signal as follows

$$\widehat{R}_x\left(e^{i2\pi\frac{\nu}{n}}\right) = \frac{1}{n}\left|X\left(e^{i2\pi\frac{\nu}{n}}\right)\right|^2.$$

Since all frequencies $f_k$ are multiple of $\frac{1}{n}$, the discrete spectrum $R_x\left(e^{i2\pi\frac{\nu}{n}}\right) \triangleq \mathbb{E}\left[\widehat{R}_x\left(e^{i2\pi\frac{\nu}{n}}\right)\right]$ is such that

$$R_x\left(e^{i2\pi\frac{\nu}{n}}\right) = \sigma^2 + \sum_{k=0}^{K-1} a_k^2\, \mathbf{1}_{\left\{e^{i2\pi\frac{\nu}{n}} = z_k\right\}}. \tag{5.17}$$

Furthermore, since all frequencies $f_k$ are multiple of $\frac{1}{l}$, we can verify that

$$\boldsymbol{P}(t) = \mathrm{diag}(a_0^2, \ldots, a_{(K-1)}^2).$$

Therefore, $\boldsymbol{R}_{xx}(t) = \sigma^2\boldsymbol{I}_n + \sum_{k=0}^{K-1} a_k^2 \boldsymbol{v}(z_k)\,\boldsymbol{v}(z_k)^H$. It is then assumed without loss of generality that the $a_k$ are sorted in decreasing order. However, the family $\{\boldsymbol{v}(z_k)\}_{k=0\ldots K-1}$ is orthogonal. Therefore $\forall m \in \{0 \ldots K-1\}$, $\boldsymbol{R}_{xx}(t)\boldsymbol{v}(z_k) = (a_k^2 + \sigma^2)\boldsymbol{v}\boldsymbol{b}(z_k)$. Thus, $\{\boldsymbol{v}(z_k)\}_{k=0\ldots K-1}$ forms a family of eigenvectors of matrix $\boldsymbol{R}_{xx}(t)$ , associated to the eigenvalues $\{a_k^2 + \sigma^2\}_{k=0\ldots K-1}$. By completing this family with the other vectors of the $n$-th order Fourier basis, we obtain a basis of eigenvectors of matrix $\boldsymbol{R}_{xx}(t)$ (all other eigenvectors being associated with the same eigenvalue $\sigma^2$).

Consequently, the spectrum of the eigenvalues of matrix $\boldsymbol{R}_{xx}(t)$ matches the discrete spectrum given in equation (5.17) (whose periodogram forms an estimator ).

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

### 3.2.4 MUltiple SIgnal Characterization (MUSIC)

The MUSIC method, developed by R. O. Schmidt [37], is based on the following remark: the poles $\{z_k\}_{k=0...K-1}$ are the only solutions of equation

$$\|\boldsymbol{W}_\perp(t)^H \boldsymbol{v}(z)\|^2 = 0 \tag{5.18}$$

where $\boldsymbol{v}(z) = [1, z, , \ldots, z^{n-1}]^\top$. Indeed, $z$ is solution if and only if $\boldsymbol{v}(z) \in \text{span}(\boldsymbol{W}(t)) = \text{span}(\boldsymbol{V}^n)$. So every pole $z_k$ is a solution, and there can be no other because otherwise the signal subspace would be of dimension strictly larger than $K$. So the *root-MUSIC* [27] method consists of the following stages:

- calculate and diagonalize matrix $\widehat{\boldsymbol{R}}_{xx}(t)$;
- deduce a basis of the noise subspace $\boldsymbol{W}_\perp(t)$;
- extract the roots of equation (5.18).

In the particular case where the noise subspace has dimension 1, it is equivalent to the Pisarenko method presented in section 3.1.3.

In practice, real signals do not strictly correspond to the model, and equation (5.18) is not rigorously verified. This is why the *spectral-MUSIC* [26] method rather consists in searching for the $K$ highest peaks of function $\widehat{S}(z) = \frac{1}{\|\boldsymbol{W}_\perp^H \boldsymbol{v}(z)\|^2}$. The *spectral-MUSIC* method is illustrated in figure I.3, where it is applied to a piano note.

The ESPRIT method, presented below, avoids the optimization of function $\widehat{S}(z)$, or the resolution of equation (5.18), and provides the values of the complex poles in a more direct way.

### 3.2.5 Estimation of Signal Parameters via Rotational Invariance Techniques

The ESPRIT [29] method is based on a particular property of the signal subspace: the rotational invariance. Let $\boldsymbol{V}^n_\downarrow$ be the matrix of dimensions $(n-1)\times K$ which contains the first $n-1$ rows of $\boldsymbol{V}^n$, and $\boldsymbol{V}^n_\uparrow$ the matrix of dimensions $(n-1)\times K$ which contains the last $n-1$ rows of $\boldsymbol{V}^n$. Similarly, let $\boldsymbol{W}(t)_\downarrow$ be the matrix of dimensions $(n-1)\times K$ which contains the first $n-1$ rows of $\boldsymbol{W}(t)$, and $\boldsymbol{W}(t)_\uparrow$ the matrix of dimensions $(n-1)\times K$ which contains the $n-1$ last rows of $\boldsymbol{W}(t)$. Then we have

$$\boxed{\boldsymbol{V}^n_\uparrow = \boldsymbol{V}^n_\downarrow \boldsymbol{D}} \tag{5.19}$$

where $\boldsymbol{D} = \text{diag}(z_0, \ldots, z_{(K-1)})$. Now the columns of $\boldsymbol{V}^n$ and those of $\boldsymbol{W}(t)$ form two bases of the same vector space of dimension $K$. Thus, there is an invertible matrix $\boldsymbol{G}(t)$ of dimension $K\times K$ such that

$$\boldsymbol{V}^n = \boldsymbol{W}(t)\,\boldsymbol{G}(t) \tag{5.20}$$

where $\boldsymbol{G}(t)$ is defined as the transition matrix from the first basis to the second one. By substituting equation (5.20) into equation (5.19), we show that

$$\boxed{\boldsymbol{W}(t)_\uparrow = \boldsymbol{W}(t)_\downarrow\,\boldsymbol{\Phi}(t)} \tag{5.21}$$

where $\boldsymbol{\Phi}(t)$, called *spectral matrix*, is defined by its EVD:

$$\boxed{\boldsymbol{\Phi}(t) = \boldsymbol{G}(t)\,\boldsymbol{D}\,\boldsymbol{G}(t)^{-1}.} \tag{5.22}$$

In particular, the eigenvalues of $\boldsymbol{\Phi}(t)$ are the poles $\{z_k\}_{k=0...K-1}$.

By multiplying equation (5.21) on the left by $\boldsymbol{W}(t)^H_\downarrow$, we get

$$\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\uparrow = \boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow\,\boldsymbol{\Phi}(t). \tag{5.23}$$

However, if $\text{rank}(\boldsymbol{W}(t)_\downarrow) = K$, matrix $\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow$ is invertible. Indeed, we trivially note that $\forall \boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow \boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \boldsymbol{W}(t)_\downarrow \boldsymbol{x} = 0$. So $\dim(\ker(\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow)) = \dim(\ker(\boldsymbol{W}(t)_\downarrow))$. The rank-nullity theorem allows us to conclude that $\text{rank}(\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow) = \text{rank}(\boldsymbol{W}(t)_\downarrow) = K$, so $\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow$ is invertible. Therefore equation (5.23) implies $\boldsymbol{\Phi}(t) = \left(\boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\downarrow\right)^{-1} \boldsymbol{W}(t)^H_\downarrow \boldsymbol{W}(t)_\uparrow$.

Finally, the ESPRIT algorithm is composed of four steps:

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

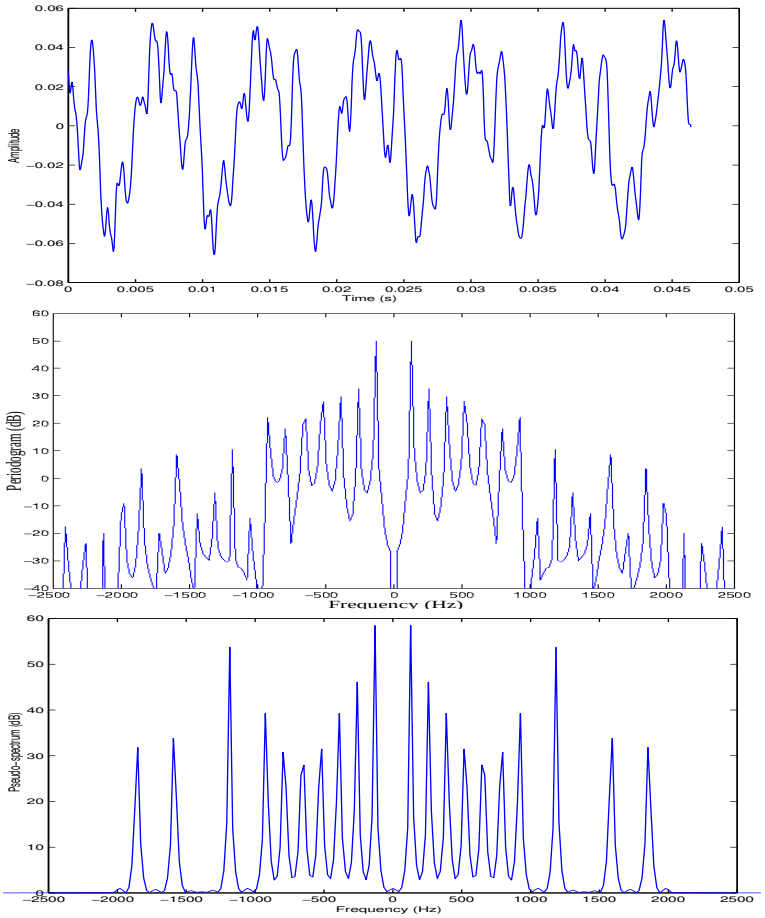



Figure 5.3: Waveform ($t \mapsto x(t)$), periodogram ($f \mapsto 20\log_{10}|S(e^{i2\pi f})|$) and pseudo-spectrum ($f \mapsto 20\log_{10}|\widehat{S}(e^{i2\pi f})|$ with $K = 20$ and $n = 256$) of a piano note

- calculate and diagonalize matrix $\widehat{\boldsymbol{R}}_{xx}(t)$;
- deduce a basis of the signal subspace $\boldsymbol{W}(t)$;
- extract from $\boldsymbol{W}(t)$ matrices $\boldsymbol{W}(t)_{\downarrow}$ and $\boldsymbol{W}(t)_{\uparrow}$;
- compute the spectral matrix $\boldsymbol{\Phi}(t) = \left(\boldsymbol{W}(t)_{\downarrow}^H \boldsymbol{W}(t)_{\downarrow}\right)^{-1} \boldsymbol{W}(t)_{\downarrow}^H \boldsymbol{W}(t)_{\uparrow}$
- diagonalize $\boldsymbol{\Phi}(t)$ and deduce the estimated poles.

Theoretical and experimental studies have shown that the ESPRIT method is the most efficient of the HR methods presented above. (*cf.* section 5.2).

# 4 Estimation of the other parameters

The high resolution methods exposed in the previous sections estimate only the poles $z_k$. We are now interested in the estimation of the other model parameters.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

## 4.1 Estimation of the modeling order

Until now, the order of the ESM model was assumed to be known, which is generally not the case in practice. Many methods have been proposed in the literature for estimating the number of sinusoids present in white noise. The most classic ones are the maximum likelihood method [38] and the *Information Theoretic Criteria* (ITC) [39], among which the *Akaike Information Criterion* (AIC) criterion [40] and the *Minimum Description Length* (MDL) criterion by Schwartz [41] and Rissanen [42]. Another technique in the context of ITC is the *Efficient Detection Criteria* (EDC) criterion [43], which is also robust to a multiplicative white noise [44]. These various ITC criteria are based on the similarity of the eigenvalues in the noise subspace, and not on the existence of a gap between the signal and noise subspaces [45]. A criterion for selecting the modeling order based on this gap, formulated in terms of *maximally stable* decomposition, was developed in [46]. Other approaches are based on Wishart matrices [47] and on the cross validation method [48].

However, in the case where the noise is colored, all these methods tend to overestimate the order of the model. Thus, specific methods have been designed to deal with the case of colored noise, among which new ITC criteria [49, 50], a technique based on a model of noise ACF of finite support [51], and a maximum a posteriori criterion [52].

Among all these methods, we present here the most classic ones, namely the three main ITC criteria: AIC,MDL and EDC (which is a robust generalization of AIC and MDL). These methods consist in minimizing a cost function composed of a first common term and a second term which forms a penalizing factor:

$$\boxed{\mathrm{ITC}(p) = -(n-p)\,l\,\ln\left(\frac{\left(\prod_{q=p+1}^{n}\sigma_q^2\right)^{\frac{1}{n-p}}}{\frac{1}{n-p}\sum_{q=p+1}^{n}\sigma_q^2}\right) + p\,(2n-p)\,C(l)}$$

where the scalars $\sigma_q^2$ are the eigenvalues of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ sorted in decreasing order, and $C(l)$ is a function of variable $l$. The AIC criterion is defined by setting $C(l) = 1$, and the MDL criterion is defined by setting $C(l) = \frac{1}{2}\ln(l)$ . The EDC criteria are obtained for all functions $l \mapsto C(l)$ such that $\lim_{l\to+\infty} \frac{C(l)}{l} = 0$ and $\lim_{l\to+\infty} \frac{C(l)}{\ln(\ln(l))} = +\infty$. These criteria lead to maximizing the ratio of the geometric mean of the eigenvalues of the noise subspace to their arithmetic mean. However this ratio is maximum and equal to 1 when all these eigenvalues are equal; it therefore measures the whiteness of the noise (in theory the eigenvalues are all equal to $\sigma^2$). The penalty term $C(l)$ avoids overestimating $p$. In practice, these methods are relatively satisfactory when processing signals that fit well the signal model, but their performance collapses when this model is less well fitted, in particular when the noise is colored.

## 4.2 Estimation of amplitudes, phases and standard deviation of noise

The maximum likelihood principle developed in section 2.1 suggests using the least squares method to estimate the complex amplitudes (*cf.* equation (5.5)):

$$\boldsymbol{\alpha}(t) = \boldsymbol{V}^{N\dagger}\boldsymbol{x}(t),$$

from which $a_k = |\alpha_k|$ and $\phi_k = \arg(\alpha_k)$ are deduced. Remember that according to the Gauss-Markov theorem, the least squares estimator is an unbiased linear estimator, with minimum variance among all unbiased linear estimators, since the additive noise is white. In the case where the additive noise is colored, the optimal estimator is obtained by the weighted least squares method (we can refer to [53] for detailed information on the estimation of amplitudes by the weighted least squares method) .

Finally, the maximum likelihood principle suggests estimating the standard deviation by calculating the power of the residual (*cf.* equation (5.4)):

$$\sigma^2 = \frac{1}{N}\left\|\boldsymbol{x}(t) - \boldsymbol{V}^N\boldsymbol{\alpha}(t)\right\|^2 .$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# 5 Performance of the estimators

## 5.1 Cramer-Rao bound

The Cramér-Rao bound is a fundamental tool in probability theory, because it makes it possible to analyze the performance of an estimator, by comparing its variance to an optimal value, which in a way acts as a quality benchmark. In the particular case of the ESM signal model, a study of the Cramér-Rao bound was proposed in [35]. The general Cramér-Rao bound theorem is summarized below (*cf.* [12]). It is based on the assumption of a regular statistical model.

**Definition 18** (Regular statistical model)**.** *Let us consider a statistical model dominated by a measure $\mu$ and parameterized by $\boldsymbol{\theta} \in \Theta$, where $\Theta$ is an open set of $\mathbb{R}^q$. Let $\boldsymbol{x}$ denote the random vector of dimension $N$. Then the parameterization is said to be regular if the following conditions are satisfied:*

1. *the probability density $p(\boldsymbol{x};\, \boldsymbol{\theta})$ is continuously differentiable, $\mu$-almost everywhere, with respect to $\boldsymbol{\theta}$.*
2. *the Fisher information matrix*

$$\boldsymbol{F}(\boldsymbol{\theta}) \triangleq \int_H \boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta})\, \boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta})^\top\, p(\boldsymbol{x};\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}$$

*defined from the score function $\boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta}) \triangleq \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\, \boldsymbol{\theta})\, \mathbf{1}_{p(\boldsymbol{x};\, \boldsymbol{\theta})>0}$ is positive definite for any value of parameter $\boldsymbol{\theta}$, and continuous with respect to $\boldsymbol{\theta}$.*

**Theorem 33** (Cramér-Rao bound)**.** *Let us consider a regular statistical model parameterized by $\boldsymbol{\theta} \in \Theta$. Let $\widehat{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}$ ($\forall \boldsymbol{\theta} \in \Theta$, $\mathbb{E}_{\boldsymbol{\theta}}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$). Then the dispersion matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \triangleq \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top\right]$ is such that matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) - \boldsymbol{F}(\boldsymbol{\theta})^{-1}$ is positive semidefinite.*

In particular, the diagonal entries of matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) - \boldsymbol{F}(\boldsymbol{\theta})^{-1}$ are non-negative. Consequently, the variances of the coefficients of $\widehat{\boldsymbol{\theta}}$ are greater than the diagonal entries of matrix $\boldsymbol{F}(\boldsymbol{\theta})^{-1}$. Thus the Cramér-Rao estimation bounds for all the scalar parameters are obtained in three stages:

- calculation of the Fisher information matrix;
- inversion of this matrix;
- extraction of its diagonal entries.

As mentioned in section 2.1, the vector $\boldsymbol{x}(t)$ containing the $N$ samples of the observed signal is a Gaussian random vector of expected value $\boldsymbol{s}(t)$ and covariance matrix $\boldsymbol{R}_{bb}$. Below, the dependence of $\boldsymbol{s}(t)$ and $\boldsymbol{R}_{bb}$ on the parameters of the model will be mentioned explicitly. On the other hand, in order to simplify the notation, we will omit the dependence of $\boldsymbol{s}(t)$ with respect to time. It is known that the Fisher information matrix of a Gaussian random vector is expressed simply as a function of the model parameters, as shown in the following proposition [12, pp. 525].

**Proposition 34** (Fisher's information matrix for a Gaussian density)**.** *For a family of complex Gaussian probability laws of covariance matrix $\boldsymbol{R}_{bb}(\boldsymbol{\theta})$ and of mean $\boldsymbol{s}(\boldsymbol{\theta})$ , where $\boldsymbol{R}_{bb} \in C^1(\Theta,\, \mathbb{C}^{N\times N})$ and $\boldsymbol{s} \in C^1(\Theta,\, \mathbb{C}^N)$, the entries of the Fisher information matrix $\left\{\boldsymbol{F}_{(i,j)}(\boldsymbol{\theta})\right\}_{1\leq i,\, j\leq k}$ are given by the extended Bangs-Slepian formula:*

$$\boxed{\boldsymbol{F}_{(i,j)}(\boldsymbol{\theta}) = \mathrm{trace}\left(\boldsymbol{R}_{bb}^{-1}\, \frac{\partial \boldsymbol{R}_{bb}(\boldsymbol{\theta})}{\partial \theta_i}\, \boldsymbol{R}_{bb}^{-1}\, \frac{\partial \boldsymbol{R}_{bb}(\boldsymbol{\theta})}{\partial \theta_j}\right) + 2\mathcal{R}e\left(\frac{\partial \boldsymbol{s}(\boldsymbol{\theta})}{\partial \theta_i}^H\, \boldsymbol{R}_{bb}^{-1}\, \frac{\partial \boldsymbol{s}(\boldsymbol{\theta})}{\partial \theta_j}\right).} \tag{5.24}$$

By applying formula (5.24) to the ESM model, we obtain a closed-form expression of the Fisher information matrix. We deduce the following theorem, proved in [35]:

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Contexte académique } sans modifications
*Voir Page 98*

**Proposition 35.** *The Cramér-Rao bounds for parameters* $(\phi_k, \delta_k, f_k)$ *are independent of* $a_{k'}$ *for all* $k' \neq k$*, but proportional to* $\frac{1}{a_k^2}$*. The bound for parameter* $a_k$ *is independent of all* $a_{k'}$*. Finally, the bounds for all parameters are independent of all the phases* $\phi_{k'}$*, and are unchanged by a translation of the set of frequencies* $f_{k'}$*.*

In addition, the Cramér-Rao bounds can be calculated in closed-form under certain assumptions, as was done in [54].

**Proposition 36.** *Suppose that all the damping factors are zero, and let us make N tend towards* $+\infty$*. Then the Cramér-Rao bounds for the parameters of the ESM model admit the following first-order expansions:*

- $\mathrm{CRB}\{\sigma\} = \frac{\sigma^2}{4N} + O\left(\frac{1}{N^2}\right)$;
- $\mathrm{CRB}\{f_k\} = \frac{6\sigma^2}{4\pi^2 N^3 a_k^2} + O\left(\frac{1}{N^4}\right)$;
- $\mathrm{CRB}\{a_k\} = \frac{2\sigma^2}{N} + O\left(\frac{1}{N^2}\right)$;
- $\mathrm{CRB}\{\phi_k\} = \frac{2\sigma^2}{N a_k^2} + O\left(\frac{1}{N^2}\right)$.

We note in particular that the Cramér-Rao bounds related to the frequencies $f_k$ are of order $\frac{1}{N^3}$, which is unusual in parametric estimation. Furthermore, it is known that the maximum likelihood principle provides asymptotically efficient estimators [12]. Thus, the variances of the estimators given in section 2.1 are asymptotically equivalent to the Cramér-Rao bounds given in proposition 36. The case of HR methods is discussed below.

## 5.2 Performance of HR methods

The performance of an estimator is generally expressed in terms of bias and variance. It is also possible to measure its efficiency, defined as the ratio between its variance and the Cramér-Rao bound. In particular, an estimator is said to be efficient if its efficiency is equal to 1.

In the case of HR methods, calculating the bias and variance in closed-form unfortunately turns out to be impossible, because the extraction of the roots of a polynomial, or of the eigenvalues of a matrix, induces a complex relationship between the statistics of the signal and those of the estimators. However, asymptotic results have been obtained thanks to the perturbation theory. These results are based either on the hypothesis $N \to +\infty$ (in the case where all poles are on the unit circle), or on the hypothesis of a high *Signal to Noise Ratio* (SNR) (SNR $\to +\infty$). Under each of these two hypotheses, it has been shown that all HR methods presented in this chapter are unbiased. Furthermore, under the assumption $N \to +\infty$, the variances of the Prony and Pisarenko methods were calculated in [55], and those of MUSIC and ESPRIT in [56]. Under the hypothesis SNR$\to +\infty$, the variance of Prony's method was calculated in [57], that of MUSIC in [58], and that of ESPRIT in [59, 58].

The mathematical developments proposed in all these articles are quite complex, and are strongly related to the estimation method considered, this is why they are not reproduced as part of this document. Only the main results are summarized here. First of all, it has been proved in [57, 55] that the Prony and Pisarenko methods are very inefficient, in the statistical sense: their variances are much greater than the Cramér-Rao bounds. In addition, they increase faster than the Cramér-Rao bounds when the SNR decreases. On the other hand, the MUSIC and ESPRIT methods have an asymptotic efficiency close to 1. More precisely, it has been proved in [56, 58] (in the case of unmodulated sinusoids) that these two methods achieve almost identical performances, but ESPRIT is slightly better than MUSIC. The study carried out in [59] (in the more general case of exponentially modulated sinusoids) goes in the same direction: ESPRIT is less sensitive to noise than MUSIC.

# 6 Conclusion

In this chapter, we have shown that the estimation of frequencies and damping factors by the maximum likelihood method leads to a difficult optimization problem. When all the poles of the signal are on the unit circle, it can be approximated by detecting the $K$ main peaks of the periodogram. This result is only valid when the length of

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

the observation window is sufficiently large compared to the inverse of the smallest frequency difference between neighboring poles. The main interest of HR methods is that they overcome this limit of Fourier analysis in terms of spectral resolution. The first methods of this family, proposed by Prony and Pisarenko, are based on the linear recurrence equations which characterize the signal model. On the other hand, more modern techniques, including the MUSIC and ESPRIT methods, rely on the decomposition of the data space into two eigensubspaces of the covariance matrix, called signal subspace and noise subspace. The statistical study of these various estimation techniques has shown that the ESPRIT method is the most efficient. The amplitudes and phases of the complex exponentials can then be estimated by the least squares method.

# 7 Appendices

## 7.1 Constrained optimization

Let $V$ be a Hilbert space and $F$ be a closed subset of $V$ defined by $F = \{\boldsymbol{y} \in V / f_1(\boldsymbol{y}) = 0 \dots f_M(\boldsymbol{y}) = 0\}$, where functions $\{f_m\}_{m=1\dots M}$ are continuous from $V$ to $\mathbb{R}$. Let $J$ be a real function on $V$ and $\boldsymbol{p}$ be a local minimum of $J$ over $F$. We also assume that functions $J$ and $\{f_m\}_{m=1\dots M}$ are differentiable at $\boldsymbol{p}$.

**Theorem 37** (Lagrange multipliers)**.** *There is $\mu_0, \dots \mu_M \in \mathbb{R}$ not all zero such that $\mu_0 J'(\boldsymbol{p}) + \sum_{m=1}^{M} \mu_m f_m'(\boldsymbol{p}) = 0$. If in addition the vectors $\{f_m'(\boldsymbol{p})\}_{m=1\dots M}$ are linearly independent, then there are coefficients $\lambda_1 \dots \lambda_M \in \mathbb{R}$ called Lagrange multipliers such that $J'(\boldsymbol{p}) + \sum_{m=1}^{M} \lambda_m f_m'(\boldsymbol{p}) = 0$.*

## 7.2 Vandermonde matrices

**Definition 19** (Vandermonde matrix)**.** *Let $K \in \mathbb{N}^*$ and $z_0, \dots, z_{K-1} \in \mathbb{C}$. We call Vandermonde matrix a matrix $\boldsymbol{V}$ of dimension $K \times K$ of the form*

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ {z_0}^{K-1} & {z_1}^{K-1} & \dots & {z_{K-1}}^{K-1} \end{bmatrix}.$$

**Proposition 38** (Vandermonde determinant)**.** *The determinant of the Vandermonde matrix is*

$$\det(\boldsymbol{V}) = \prod_{0 \le k_1 < k_2 \le K-1} (z_{k_2} - z_{k_1}).$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Contexte académique } sans modifications
*Voir Page 98*

# Chapter 6

# 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. This chapter was written by François Roueff.

## 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 $\sigma$-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:

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

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

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

And, as for standard expectation, if $\mathbf{X}$ is a random vector, $\mathbb{E}[\mathbf{X}|\mathcal{G}]$ 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 8.1).

**Proposition 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*

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

*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. 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*

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

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

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

*and*

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

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# 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 [19, 20] 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 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 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 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 $t \geq 1$,*

$$\mathbf{X}_t = \Phi_t \mathbf{X}_{t-1} + \mathrm{A}_t \mathbf{u}_t + \mathbf{W}_t \,, \tag{6.1}$$
$$\mathbf{Y}_t = \Psi_t \mathbf{X}_t + \mathrm{B}_t \mathbf{u}_t + \mathbf{V}_t \,, \tag{6.2}$$

*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 $\mathrm{A}_t$ 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) $\Psi_t$ is a $q \times p$* measurement *or* observation matrix *for all $t \geq 1$,*

*(v) The matrix $\mathrm{B}_t$ 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 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 3.3.

**Remark 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 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* (6.1) 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* (6.2) 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 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 $t \geq 1$,*

$$Z_{t+1} = Z_t + \beta + W_t = Z_1 + \beta t + W_1 + \cdots + W_t \,, t \geq 0 \,.$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

*When $\sigma$ 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 6.1.*

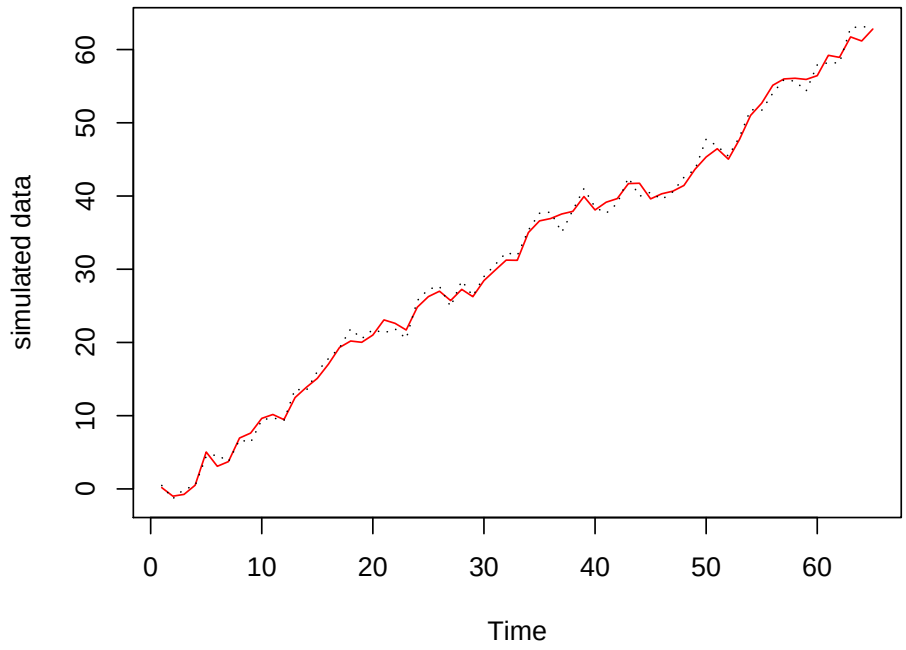



Figure 6.1: Simulated random trend (plain red line) and its observation with additive noise (dotted black line).

**Example 2.2** (Climatology data)**.** *Figure 6.2 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* $\mathrm{B}_t = 0$.

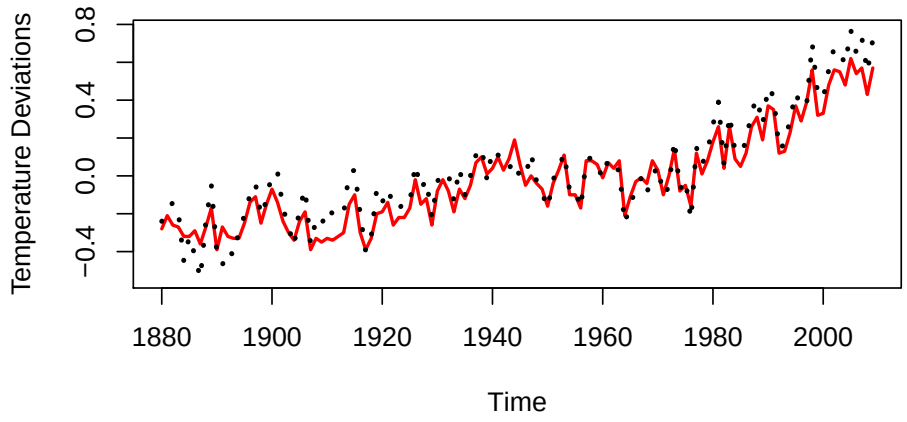



Figure 6.2: Annual global temperature deviation series, measured in degrees centigrade, 1880–2009.

Dynamic linear models allow us to provide a quite general framework for denoising and forecasting a Gaussian process, or/and estimating its parameters. In (6.1) and (6.2), 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 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 4 page 30, 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 6.3. We easily compute*

$$\gamma_y(0) = \text{Var}\,(Y_t) = \text{Var}\,(X_t + V_t) = \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2, \tag{6.3}$$

*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 8.3). We will provide a general view on the relationships between DLMs and stationary ARMA processes in Section 6.*

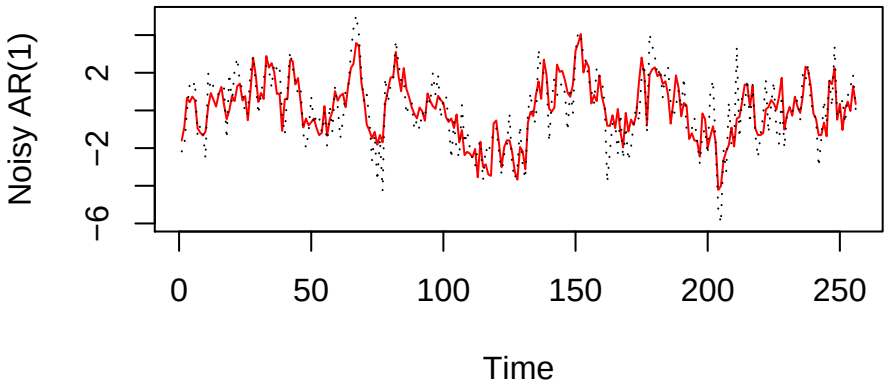



Figure 6.3: Simulated AR(1) process (solid red) and a noisy observation of it (dotted black).

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

$$\mathbf{X}_{t|s} \overset{\text{def}}{=} \mathbb{E}\,[\mathbf{X}_t | \mathbf{Y}_{1:s}] \tag{6.4}$$

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

$$\mathbf{\Sigma}_{t|s} \overset{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_t - \mathbf{X}_{t|s})(\mathbf{X}_t - \mathbf{X}_{t|s})^T\right] = \text{Cov}(\mathbf{X}_t - \mathbf{X}_{t|s})\,, \tag{6.5}$$

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

Three different situations are generally distinguished.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

a- It is called a *forecasting* or prediction problem if $s < t$.

b- It is called a *filtering* problem if $s = t$.

c- It is called a *smoothing* problem if $s > t$.

Interestingly, these tasks are very much related to the computation of the likelihood for estimating the unknown parameters of the models, see Section 7.

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:

**Data:** Parameters $Q$, $R$ and $\mathrm{A}_t$, $\mathrm{B}_t$, $\Psi_t$ for $t = 1, \dots, n$, initial conditions $\boldsymbol{\mu}$ and $\Sigma_0$, observations $\mathbf{Y}_t$ and exogenous input series $\mathbf{u}_t$, for $t = 1, \dots, n$.
**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 $t = 1, \dots, n$.
Initialization: set $\mathbf{X}_{0|0} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{0|0} = \Sigma_0$.
**for** $t = 1, 2, \dots, n$ **do**
Compute in this order

$$\mathbf{X}_{t|t-1} = \Phi_t \mathbf{X}_{t-1|t-1} + \mathrm{A}_t \mathbf{u}_t, \tag{6.6}$$
$$\boldsymbol{\Sigma}_{t|t-1} = \Phi_t \boldsymbol{\Sigma}_{t-1|t-1} \Phi_t^T + Q, \tag{6.7}$$
$$K_t = \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T [\Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R]^{-1}. \tag{6.8}$$
$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + K_t(\mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t), \tag{6.9}$$
$$\boldsymbol{\Sigma}_{t|t} = [I - K_t \Psi_t] \boldsymbol{\Sigma}_{t|t-1}. \tag{6.10}$$

**end**

**Algorithm 1:** Kalman filter algorithm.

**Proposition 3.1** (Kalman Filter)**.** *Algorithm 1 holds for the state-space model satisfying Assumption 2.1, provided that $\Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R$ are invertible matrices for $t = 1, \dots, n$.*

The matrix $K_t$ defined in (6.8) is called the *Kalman gain matrix*.

For proving Proposition 3.1, we will introduce the following definition

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

and show the folowing useful formula

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t \mathbf{X}_{t|t-1} - \mathrm{B}_t \mathbf{u}_t, \tag{6.11}$$
$$\Gamma_t = \mathrm{Cov}(\Psi_t(\mathbf{X}_t - \mathbf{X}_{t|t-1}) + \mathbf{V}_t) = \Psi_t \boldsymbol{\Sigma}_{t|t-1} \Psi_t^T + R \tag{6.12}$$

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

*Proof of Proposition 3.1.* By Assumption 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 (6.1), this implies that, for all $t > s$

$$\begin{aligned} \mathbf{X}_{t|s} &= \mathbb{E}\left[\mathbf{X}_t | \mathbf{Y}_{1:s}\right] \\ &= \mathbb{E}\left[\Phi_t \mathbf{X}_{t-1} + \mathrm{A}_t \mathbf{u}_t + \mathbf{W}_t | \mathbf{Y}_{1:s}\right] \\ &= \Phi_t \mathbf{X}_{t-1|s} + \mathrm{A}_t \mathbf{u}_t\ , \end{aligned} \tag{6.13}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

and, moreover,

$$\begin{aligned}\boldsymbol{\Sigma}_{t|s} &= \mathrm{Cov}\left(\mathbf{X}_t - \mathbf{X}_{t|s}\right) \\ &= \mathrm{Cov}\left(\Phi_t(\mathbf{X}_{t-1} - \mathbf{X}_{t-1|s}) + \mathbf{W}_t\right) \\ &= \Phi_t \boldsymbol{\Sigma}_{t-1|s} \Phi_t^T + Q \ . \end{aligned} \tag{6.14}$$

which gives (6.6) and (6.7).

Next, we show (6.9). By definition of the innovation process, the $\sigma$-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 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 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 (6.2), 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

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

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

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - K_t \Psi_t \boldsymbol{\Sigma}_{t|t-1} \ ,$$

with

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

That is, we have shown (6.8), (6.9) and (6.10) and the proof is concluded. □

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

Let us consider the forecasting and smoothing problems, that is the computation of $\mathbf{X}_{t|n}$ for $t > n$ and $t = 1, \dots, n-1$, successively. These algorithms complete Algorithm 1 in the sense that in practice one can use them after having first applied Algorithm 1.

**Data:** A forecasting lag $h$, parameters $Q$ and $\mathrm{A}_t$ 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

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

**end**

**Algorithm 2:** Kalman forecasting algorithm.

**Data:** Parameters $\Phi_t$ 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

$$J_{t-1} = \mathbf{\Sigma}_{t-1|t-1}\Phi_t^T\mathbf{\Sigma}_{t|t-1}^{-1}\ , \tag{6.15}$$

$$\mathbf{X}_{t-1|n} = \mathbf{X}_{t-1|t-1} + J_{t-1}\left(\mathbf{X}_{t|n} - \mathbf{X}_{t|t-1}\right), \tag{6.16}$$

$$\mathbf{\Sigma}_{t-1|n} = \mathbf{\Sigma}_{t-1|t-1} + J_{t-1}\left(\mathbf{\Sigma}_{t|n} - \mathbf{\Sigma}_{t|t-1}\right)J_{t-1}^T\ . \tag{6.17}$$

**end**

**Algorithm 3:** Rauch-Tung-Striebel smoother algorithm.

**Proposition 3.2.** *Algorithm 2 and Algorithm 3 hold for the state-space model satisfying Assumption 2.1, provided that (only for Algorithm 3) $\mathbf{\Sigma}_{t|t-1}$ is an invertible matrix for $t = 2, \dots, n$.*

*Proof.* Algorithm 2 directly follows from (6.13) and (6.14).

We now show that Algorithm 3 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

$$\mathbb{E}\left[\mathbf{X}_{t-1}\middle|\mathbf{Y}_{1:n}\right] = \mathbb{E}\left[\widetilde{\mathbf{X}}_{t-1}\middle|\mathbf{Y}_{1:n}\right]\ , \tag{6.18}$$

where

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

since $\mathbf{V}_{t:n}, \mathbf{W}_{t+1:n}$ are independent of all other variables appearing in this formula. Using the Gaussian assumption and the fact that $\mathbf{Y}_{1:t-1}$ and $\mathbf{X}_t - \mathbf{X}_{t|t-1}$ are uncorrelated, we get

$$\widetilde{\mathbf{X}}_{t-1} = \mathbf{X}_{t-1|t-1} + J_{t-1}(\mathbf{X}_t - \mathbf{X}_{t|t-1}), \tag{6.19}$$

and

$$\mathrm{Cov}\left(\mathbf{X}_{t-1} - \widetilde{\mathbf{X}}_{t-1}\right) = \mathbf{\Sigma}_{t-1|t-1} - J_{t-1}\mathbf{\Sigma}_{t|t-1}J_{t-1}^T, \tag{6.20}$$

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 (6.15). By (6.18) and (6.19), we obtain, by projecting $\widetilde{\mathbf{X}}_{t-1}$ on 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 (6.16) 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 (6.20), 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 (6.17). □

Inspecting the proofs of Proposition 3.1 and Proposition 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 3.3.** *Suppose that Assumption 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 1, Algorithm 2 and Algorithm 3 continue to hold if the definitions of* $\mathbf{X}_{s|t}$ *in (6.6) 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* Span $(\dots)$ *is understood as the space of* $\mathbb{R}^p$*-valued* $L^2$ *random variables obtained by linear transformations of* $\dots$ *and* 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

$$\mathbf{\Sigma}_{t_1,t_2|s} \overset{\text{def}}{=} \mathbb{E}\left[(\mathbf{X}_{t_1} - \mathbf{X}_{t_1|s})(\mathbf{X}_{t_2} - \mathbf{X}_{t_2|s})^T\right] \tag{6.21}$$

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

$$\begin{aligned}\mathbf{X}_{(t)} &\overset{\text{def}}{=} [\mathbf{X}_t^T \ \mathbf{X}_{t-1}^T]^T \ , \\ \mathbf{Y}_{(t)} &\overset{\text{def}}{=} [\mathbf{Y}_t^T \ \mathbf{Y}_{t-1}^T]^T \ .\end{aligned}$$

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 8.2).

**Data:** Parameters $\Psi_n$ and $\Phi_t$ 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

$$\mathbf{\Sigma}_{n,n-1|n} = (I - K_n\Psi_n)\Phi_n\mathbf{\Sigma}_{n-1|n-1}, \tag{6.22}$$

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

$$\mathbf{\Sigma}_{t-1,t-2|n} = \mathbf{\Sigma}_{t-1|t-1}J_{t-2}^T + J_{t-1}\left(\mathbf{\Sigma}_{t,t-1|n} - \Phi_t\mathbf{\Sigma}_{t-1|t-1}\right)J_{t-2}^T. \tag{6.23}$$

**end**

**Algorithm 4:** One-lag covariance algorithm.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**Remark 3.1.** *All the above algorithms (Algorithms 1, 2, 3 and 4) are* recursive *in the sense that their outputs are computed using a simple recursive set of equations. Algorithm 1 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*** $\mathbf{Y}_1, \dots, \mathbf{Y}_t$. *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 3.2.** *It is interesting to note that in the above algorithms, the covariance matrices do not depend on the observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, 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 $\mathbf{Y}_1, \dots, \mathbf{Y}_n$.*

**Example 3.1** (Noisy AR(1) (continued from Example 2.3))**.** *Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ be as in Example 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 (6.3)). To produce Figure 6.4, the Kalman smoother was computed with these true parameters from $Y_1, \dots, Y_n$ with $n = 2^8$ only the last 16 points of $Y_t$, $X_t$ and $\mathbf{X}_{t|n}$ ($t = n-15, n-14, \dots, n$) are drawn, using respectively red circles, a dotted black line and a solid green line. The dashed blue lines represent 95% confidence intervals for $\mathbf{X}_t$ obtained using that, given $\mathbf{Y}_{1:n}$, the conditional distribution of each $\mathbf{X}_t$ is $\mathcal{N}(\mathbf{X}_{t|n}, \mathbf{\Sigma}_{t|n})$.*

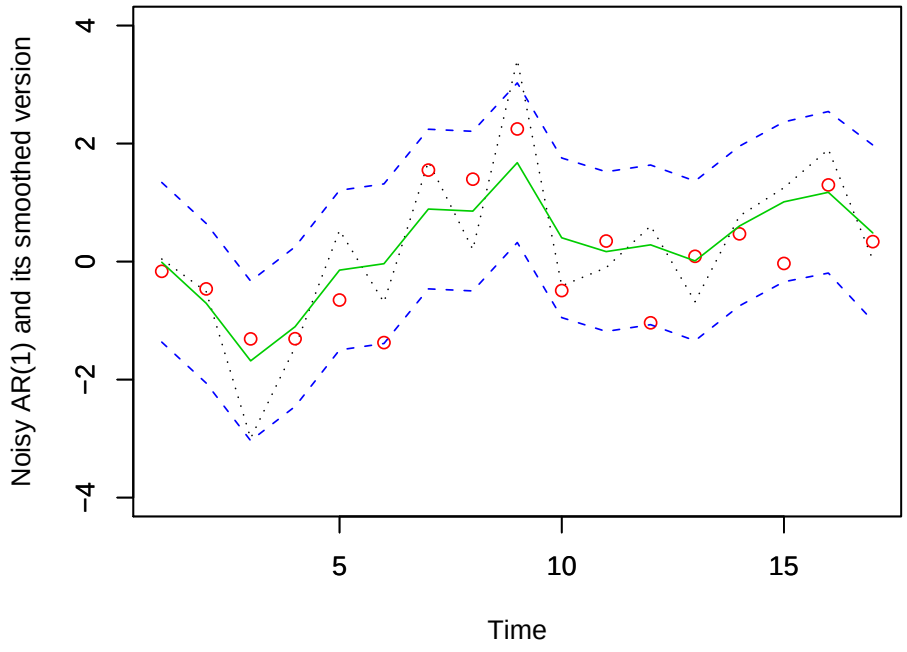



Figure 6.4: Simulated AR(1) process (red circles), a noisy observation of it (dotted black line), the smoother outputs (solid green line) and the 95% confidence intervals (between blued dashed lines).

# 4 Steady State approximations

Let us consider Assumption 2.1 in the particular case where there are no input series ($A_t = B_t = 0$) 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 1, the Kalman gain $K_t$ and the error covariance matrices $\mathbf{\Sigma}_{t|t}$ and $\mathbf{\Sigma}_{t|t-1}$ should not depend on $t$. Of course, this cannot be exactly true : these quantities correspond to state and observation variables $(\mathbf{X}_{1:t}, \mathbf{Y}_{1:t})$ whose distribution cannot be exactly the same as $((\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1})$. But it can be approximately true if the past data has a very small influence on the current ones, in other words, if the conditional distribution of $\mathbf{X}_t$ given $\mathbf{Y}_{1:t}$ is approximately the same as the conditional distribution of $\mathbf{X}_t$ given the whole past $\mathbf{Y}_{-\infty:t}$.

In practice this steady approximation of the Kalman filter is observed when $K_t \to K$ and $\mathbf{\Sigma}_{t|t-1} \to \Sigma$ as $t \to \infty$. Using the relationship between $\mathbf{\Sigma}_{t|t-1}$ and $\mathbf{\Sigma}_{t-1|t-2}$ (following (6.10) and (6.10)), we obtain that $\Sigma$ is a necessary solution to the *Ricatti equation*

$$\Sigma = \Phi[\Sigma - \Sigma\Psi^T(\Psi\Sigma\Psi^T + R)^{-1}\Psi\Sigma]\Phi^T + Q \,, \tag{6.24}$$

and following (6.8), the steady-state gain matrix reads

$$K = \Sigma\Psi^T[\Psi\Sigma\Psi^T + R]^{-1} \,.$$

The convergence of the MLE and its asymptotic normalitycan be established when $\Phi$ has eigenvalues within the open unit disk $\{z \in \mathbb{C}, |z| < 1\}$. We just refer to [21, 22] for details. Let us just briefly give a hint of why this assumption is meaningful. Iterating the state equation (6.1) in the case $\Phi_t = \Phi$ and $A_t = 0$ 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 $\Phi$ 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 1.2, one can derive the asymptotic behavior of the MLE, under appropriate assumptions of the parameterization.

# 5 Correlated Errors

Sometimes it is advantageous to use assumptions for the linear state-space model which are slightly different from Assumption 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 $\Theta$ 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 5.1.** *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 $t \geq 1$,*

$$\mathbf{X}_{t+1} = \Phi_t\mathbf{X}_t + A_{t+1}\mathbf{u}_{t+1} + \Theta_t\mathbf{W}_t \,, \tag{6.25}$$

$$\mathbf{Y}_t = \Psi_t\mathbf{X}_t + B_t\mathbf{u}_t + \mathbf{V}_t \,, \tag{6.26}$$

*where*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

*(i)* $\left(\begin{bmatrix}\mathbf{W}_t & \mathbf{V}_t\end{bmatrix}^T\right)_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} Q & S \\ S^T & R\end{bmatrix}\right)$ *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* $\mathrm{A}_t$ *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)* $\Psi_t$ *is a* $q \times p$ measurement *or* observation matrix *for all* $t \geq 1$,

*(v) The matrix* $\mathrm{B}_t$ *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 1 has to be adapted as follows.

**Data:** Parameters $Q$, $\Theta_t$, $S$, $R$ and $\mathrm{A}_t$, $\mathrm{B}_t$, $\Psi_t$ for $t = 1, \ldots, n$, initial conditions $\boldsymbol{\mu}$ and $\Sigma_0$, 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

$$\boldsymbol{\epsilon}_t = \mathbf{Y}_t - \Psi_t\mathbf{X}_{t|t-1} - \mathrm{B}_t\mathbf{u}_t \tag{6.27}$$
$$\Gamma_t = \Psi_t\boldsymbol{\Sigma}_{t|t-1}\Psi_t^T + R, \tag{6.28}$$
$$K_t = [\Phi_t\boldsymbol{\Sigma}_{t|t-1}\Psi_t^T + \Theta_t S]\Gamma_t^{-1}, \tag{6.29}$$
$$\mathbf{X}_{t+1|t} = \Phi_t\mathbf{X}_{t|t-1} + \mathrm{A}_{t+1}\mathbf{u}_{t+1} + K_t\boldsymbol{\epsilon}_t, \tag{6.30}$$
$$\boldsymbol{\Sigma}_{t+1|t} = \Phi_t\boldsymbol{\Sigma}_{t|t-1}\Phi_t^T + \Theta_t Q\Theta_t^T - K_t\Gamma_t^{-1}K_t^T, \tag{6.31}$$
$$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\Psi_t^T\Gamma_t^{-1}\boldsymbol{\epsilon}_t, \tag{6.32}$$
$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\Psi_{t+1}^T\Gamma_t^{-1}\Psi_t\boldsymbol{\Sigma}_{t|t-1}. \tag{6.33}$$

**end**

**Algorithm 5:** Kalman filter algorithm for correlated errors.

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

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

The adaptation of the proof of Proposition 3.1 to the correlated errors case is left to the reader (Exercise 8.4). The following result follows.

**Proposition 5.1** (Kalman Filter for correlated errors)**.** *Algorithm 5 applies for the state-space model satisfying Assumption 5.1, provided that* $\Psi_t\boldsymbol{\Sigma}_{t|t-1}\Psi_t^T + R$ *are invertible matrices for* $t = 1, \ldots, n$.

# 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

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \sum_{j=1}^{p}\Phi_j\mathbf{Y}_{t-j} + \sum_{k=1}^{q}\Theta_k\mathbf{V}_{t-k} + \mathbf{V}_t\ . \tag{6.34}$$

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

The observations $\mathbf{Y}_t$ are a $k$-dimensional vector process, the $\Phi$s and $\Theta$s 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 5.1, under the additional Gaussian assumption. The proof is left to the reader (Exercise 8.5).

**Proposition 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 $\Phi$ is $kp \times kp$, $\Theta$ 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*

$$\mathbf{X}_{t+1} = \Phi \mathbf{X}_t + \mathrm{A}\mathbf{u}_{t+1} + \Theta \mathbf{V}_t, \tag{6.35}$$

$$\mathbf{Y}_t = \Psi \mathbf{X}_t + \mathbf{V}_t, \tag{6.36}$$

*where $\Psi = \left[\mathbf{1}, 0, \cdots, 0\right]$ is $k \times pk$, implies the ARMAX model (6.34). If $p < q$, set $\Phi_{p+1} = \cdots = \Phi_q = 0$, and replace the value of $p$ by that of $q$ and (6.35)–(6.36) still apply. Note that the state process is $kp$-dimensional, whereas the observations are $k$-dimensional.*

**Example 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 6.1, we can write the model as*

$$X_{t+1} = \phi X_t + \beta_0 + \beta_1 t + (\theta + \phi) V_t, \tag{6.37}$$

*and*

$$Y_t = X_t + V_t. \tag{6.38}$$

**Remark 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 6.2** (Regression with autocorrelated errors)**.** *The (multivariate) regression with autocorrelated errors, is the regression model*

$$\mathbf{Y}_t = \mathrm{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t \, , \tag{6.39}$$

*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*

$$\mathbf{X}_{t+1} = \Phi \mathbf{X}_t + \Theta \mathbf{V}_t, \tag{6.40}$$

$$\mathbf{Y}_t = \Psi \mathbf{X}_t + \mathrm{B}\mathbf{u}_t + \mathbf{V}_t, \tag{6.41}$$

*where the model matrices $\Phi$, $\Theta$, and $\Psi$ are defined in Proposition 6.1.*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# 7 Likelihood of dynamic linear models

The dynamic linear model of Assumption 2.1 rely on a lot of parameters, namely $(\Psi_t)_{t\geq 1}$, $(\mathrm{A}_t)_{t\geq 1}$, $Q$, $\boldsymbol{\mu}_0$, $\Sigma_0$, $(\Phi_t)_{t\geq 1}$, $(\mathrm{B}_t)_{t\geq 1}$, 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 2.1 holds and moreover that the unknown parameters are not evolving with time. We denote by $\theta^*$ 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 $\theta$ running through a given finite dimensional parameter set $\Theta$ and with $\theta^*$ 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 $\theta^*$, the variables $\mathbf{X}_{1:n}$ are *unobserved*: one says that they are *hidden variables*.

In the following, we adapt the notation introduced in 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 (6.11) and (6.12) become

$$\boldsymbol{\epsilon}_t(\theta) = \mathbf{Y}_t - \Psi_t(\theta)\mathbf{X}_{t|t-1}(\theta) - \mathrm{B}_t(\theta)\mathbf{u}_t, \tag{6.42}$$

$$\Gamma_t(\theta) = \Psi(\theta)\boldsymbol{\Sigma}_{t|t-1}(\theta)\Psi_t(\theta)^T + R(\theta) \tag{6.43}$$

Here $\mathbf{X}_{t|t-1}(\theta)$ and $\boldsymbol{\Sigma}_{t|t-1}(\theta)$ also depend on $\theta$ since they are now functions of the parameter $\theta$ which determines the joint distribution of the hidden and observed data $\mathbf{X}_t, \mathbf{Y}_t$, $t = 1, \ldots, n$.

In the framework of a parameterized dynamic linear model, we get

$$-2\log L_n(\theta) = n\log(2\pi) + \sum_{t=1}^{n} \log \det\Gamma_t(\theta) + \sum_{t=1}^{n} \boldsymbol{\epsilon}_t(\theta)^T\Gamma_t(\theta)^{-1}\boldsymbol{\epsilon}_t(\theta)\ , \tag{6.44}$$

provided that $\Gamma_t(\theta)$ is invertible for all $t = 1, \ldots, n$ and $\theta \in \Theta$.

Observe that, for each $\theta$, the negated log likelihood $-2\log L_n(\theta)$ can thus be efficiently computed by running the Kalman filter (see Algorithm 1) and then applying (6.42), (6.43) and (6.44).

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 $\theta$.

However one needs to adapt Algorithm 1 to compute the gradient or the Hessian. A rather simple case is obtained when the $\Psi_t$s are known design matrices (that is, they do not depend on $\theta$). In this case differentiating within Algorithm 1 provides the following algorithm.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

**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 $Q$, $R$, $\mathrm{A}_t$, $\mathrm{B}_t$, $\Psi_t$ for $t = 1, \ldots, n$, $\boldsymbol{\mu}$ and $\Sigma_0$ can be evaluated at $\theta$. Functions $K_t$, $\mathbf{X}_{t|t-1}$, $\mathbf{X}_{t|t}$, $\boldsymbol{\Sigma}_{t|t-1}$, $\boldsymbol{\Sigma}_{t|t}$, $\Gamma_t$ and $\boldsymbol{\epsilon}_t$ are already computed at $\theta$ 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 $\theta$.
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 $\theta$)

$$
\begin{aligned}
\partial_i \mathbf{X}_{t|t-1} &= [\partial_i \Phi_t]\mathbf{X}_{t-1|t-1} + \Phi_t[\partial_i \mathbf{X}_{t-1|t-1}] + [\partial_i \mathrm{A}_t]\mathbf{u}_t, \\
\partial_i \boldsymbol{\Sigma}_{t|t-1} &= [\partial_i \Phi_t]\boldsymbol{\Sigma}_{t-1|t-1}\Phi_t^T + \Phi_t[\partial_i \boldsymbol{\Sigma}_{t-1|t-1}]\Phi_t^T \\
&\qquad + \Phi_t \boldsymbol{\Sigma}_{t-1|t-1}[\partial_i \Phi_t]^T + \partial_i Q, \\
\partial_i \boldsymbol{\epsilon}_t &= -\Psi_t[\partial_i \mathbf{X}_{t|t-1}] - [\partial_i \mathrm{B}_t]\mathbf{u}_t, \\
\partial_i \Gamma_t &= \Psi_t[\partial_i \boldsymbol{\Sigma}_{t|t-1}]\Psi_t^T + \partial_i R(\theta) \\
\partial_i K_t &= \left\{[\partial_i \boldsymbol{\Sigma}_{t|t-1}]\Psi_t^T - K_t[\partial_i \Gamma_t]\right\}\Gamma_t^{-1}. \\
\partial_i \mathbf{X}_{t|t} &= [\partial_i \mathbf{X}_{t|t-1}] + [\partial_i K_t]\boldsymbol{\epsilon}_t + K_t[\partial_i \boldsymbol{\epsilon}_t], \\
\boldsymbol{\Sigma}_{t|t} &= [\partial_i K_t]\Psi_t \boldsymbol{\Sigma}_{t|t-1} + [I - K_t\Psi_t][\partial_i \boldsymbol{\Sigma}_{t|t-1}].
\end{aligned}
$$

**end**

**Algorithm 6:** Kalman filter algorithm for the gradient of the likelihood.

Algorithm 1 and Algorithm 6 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 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 7.1** (Noisy AR(1) (continued from Example 2.3 and Example 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 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 6.5.*

[1] the quasi Newton procedure implemented in the `optim()` function of the R software, [23]

**Roland Badeau and François Roueff**
`roland.badeau@telecom-paris.fr`

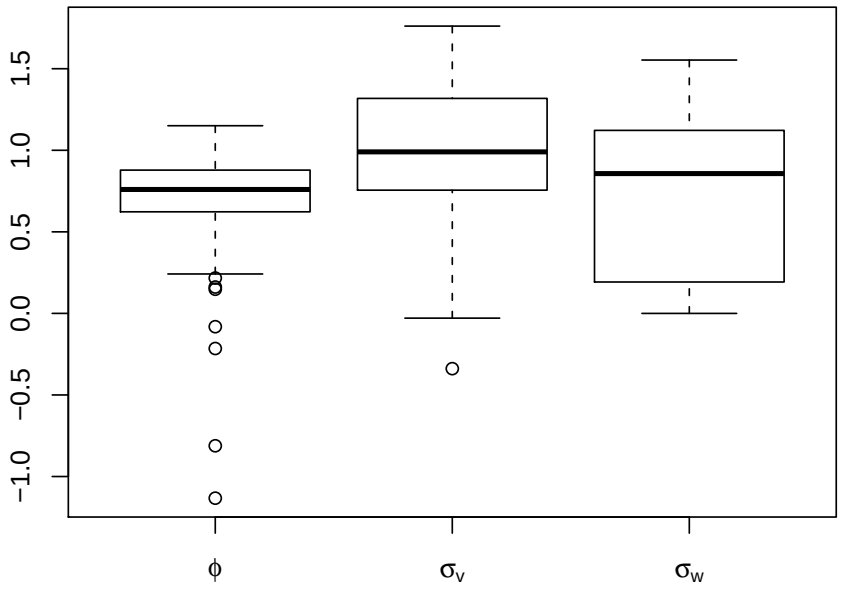



Figure 6.5: Estimation of the parameters of the noisy AR(1) model: boxplots of the estimates of $\phi$, $\sigma_v$ and $\sigma_w$ obtained from 100 Monte Carlo replications of time series of length 128. The true values are $\phi = 0.8$ and $\sigma_v = 1.0$ and $\sigma_w = 1.0$.

# 8 Exercises

**Exercise 8.1.** *Let* $\mathbf{X}$, $\mathbf{Y}$ *and* $\widehat{\mathbf{X}}$ *be as in Proposition 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*
$$\mathbb{E}\left[\boldsymbol{\epsilon}^T(\mathbf{a}' + A'\mathbf{Y})\right] = 0\,.$$
2. *Show that* $\mathbb{E}[\boldsymbol{\epsilon}] = 0$.
3. *Show that Proposition 1.1(i) holds.*
4. *Show that* $\boldsymbol{\epsilon}$ *and* $\mathbf{Y}$ *are jointly Gaussian and independent. Deduce that Proposition 1.1(ii) holds.*
5. *Show that Proposition 1.1(iii) holds.*

**Exercise 8.2.** *Show that Algorithm 4 applies under the assumptions of Proposition 3.1.*

**Exercise 8.3.** *Show that the process* $(Y_t)_{t\in\mathbb{Z}}$ *of Example 2.3 is an ARMA(1,1) process.*

**Exercise 8.4.** *Show that Algorithm 5 applies for the state-space model satisfying Assumption 5.1, provided that* $\Psi_t\boldsymbol{\Sigma}_{t|t-1}\Psi_t^T + R$ *are invertible matrices for* $t = 1,\ldots,n$.

**Exercise 8.5.** *Prove Proposition 6.1.*

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

# Bibliography


[1] L. Marple, *Digital Spectral Analysis with Applications*, Prentice Hall, Englewood Cliffs, NJ, USA, 1987.

[2] P. Stoica and R.L. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[3] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. of DAFx-02*, Hambourg, Allemagne, Sept. 2002, pp. 51–58.

[4] S. Marcos, J. Sanchez-Araujo, N. Bertaux, P. Larzabal, and P. Forster, *Les Méthodes à haute résolution : traitement d'antenne et analyse spectrale. Chapitres 4 et 5*, Hermès, Paris, France, 1998, Ouvrage collectif sous la direction de S. Marcos.

[5] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 121–132, Mar. 2004.

[6] J. Laroche, "The use of the Matrix Pencil method for the spectrum analysis of musical signals," *Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1958–1965, Oct. 1993.

[7] K. Hermus, W. Verhelst, and P. Wambacq, "Psychoacoustic modeling of audio with exponentially damped sinusoids," in *Proc. of ICASSP'02*. 2002, vol. 2, pp. 1821–1824, IEEE.

[8] C. Lambourg and A. Chaigne, "Measurements and modeling of the admittance matrix at bridge in guitars," in *Proc. of SMAC'93*, Stockholm, Suède, July 1993, pp. 449–453.

[9] G. Weinreich, "Coupled piano strings," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1474–1484, 1977.

[10] B. David, *Caractérisations acoustiques de structures vibrantes par mise en atmosphère raréfiée*, Ph.D. thesis, University of Paris VI, 1999.

[11] M. Jeanneau, P. Mouyon, and C. Pendaries, "Sintrack analysis, application to detection and estimation of flutter for flexible structures.," in *Proc. of EUSIPCO*, Ile de Rhodes, Grèce, Sept. 1998, pp. 789–792.

[12] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[13] P.H.M. Janssen and P. Stoica, "On the expectation of the product of four matrix-valued Gaussian random variables," *IEEE Trans. Autom. Control*, vol. 33, no. 9, pp. 867?–870, Sept. 1988.

[14] S. Haykin, Ed., *Nonlinear methods of spectral analysis*, Springer, Germany, 1983.

[15] P. Stoica and A. Nehorai, "On stability and root location of linear prediction models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 4, pp. 582–584, 1987.

[16] T. Söderström and P. Stoica, Eds., *System Identification*, Prentice Hall International, London, England, 1989.

[17] S. M. Kay, *Modern Spectral Estimation, Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.



**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

[18] M. Lagunas, M. Santamaria, A. Gasull, and A. Moreno, "Maximum likelihood filters in spectral estimation problems," *Signal Processing*, vol. 10, no. 1, pp. 7–18, 1986.

[19] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[20] R. E. Kalman and R. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng., Trans. ASME, Series D*, vol. 83, no. 3, pp. 95–108, 1961.

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

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

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

[24] G. M. Riche de Prony, "Essai expérimental et analytique: sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool à différentes températures," *Journal de l'école polytechnique*, vol. 1, no. 22, pp. 24–76, 1795.

[25] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical J. Royal Astron. Soc.*, vol. 33, pp. 347–366, 1973.

[26] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[27] A. J. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. of ICASSP'83*, Boston, MA, USA, 1983, pp. 336–339, IEEE.

[28] S. Y. Kung, K. S. Arun, and D. B. Rao, "State-space and singular value decomposition based approximation methods for harmonic retrieval problem," *J. of Opt. Soc. of America*, vol. 73, pp. 1799–1811, Dec. 1983.

[29] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT–A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1340–1342, Oct. 1986.

[30] R. Roy and T. Kailath, "Total least squares ESPRIT," in *Proc. of 21st Asilomar Conference on Signals, Systems, and Computers*, Nov. 1987, pp. 297–301.

[31] M. Zoltawski and D. Stavrinides, "Sensor array signal processing via a Procrustes rotations based eigenanalysis of the ESPRIT data pencil," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 6, pp. 832–861, June 1989.

[32] J. Nieuwenhuijse, R. Heusens, and Ed. F. Deprettere, "Robust exponential modeling of audio signals," in *Proc. of ICASSP'98*. May 1998, vol. 6, pp. 3581–3584, IEEE.

[33] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 6, pp. 833–840, Dec. 1982.

[34] Y. Li, K. Liu, and J. Razavilar, "A parameter estimation scheme for damped sinusoidal signals based on low-rank Hankel approximation," *IEEE Trans. Signal Process.*, vol. 45, pp. 481–486, Feb. 1997.

[35] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 5, pp. 814–824, May 1990.

[36] A-J. Van der Veen, ED. F. Deprettere, and A. L. Swindlehurst, "Subspace based signal analysis using singular value decomposition," *Proc. of IEEE*, vol. 81, no. 9, pp. 1277–1308, Sept. 1993.

[37] R. O. Schmidt, *A signal subspace approach to multiple emitter location and spectral estimation*, Ph.D. thesis, Stanford University, Stanford, Californie, USA, Nov. 1981.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

[38] G. Bienvenu and L. Kopp, "Optimality of high-resolution array processing using the eigensystem method," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1235–1245, Oct. 1983.

[39] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[40] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. of the 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds., Budapest, Hongrie, 1973, pp. 267–281, Akademia Kiado.

[41] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[42] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[43] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *Journal of Multivariate Analysis*, vol. 20, no. 1, pp. 1–25, 1986.

[44] F. Gini and F. Bordoni, "On the behavior of information theoretic criteria for model order selection of InSAR signals corrupted by multiplicative noise," *Signal Processing*, vol. 83, pp. 1047–1063, 2003.

[45] A. P. Liavas and P. A. Regalia, "On the behavior of Information Theoretic Criteria for model order selection," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1689–1695, Aug. 2001.

[46] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Blind channel approximation: effective channel order determination," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3336–3344, Dec. 1999.

[47] J. Grouffaud, P. Larzabal, and H. Clergeot, "Some properties of ordered eigenvalues of a Wishart matrix: application in detection test and model order selection," in *Proc. of ICASSP'96*. 1996, vol. 5, pp. 2465–2468, IEEE.

[48] D. Kundu and A. Mitra, "Detecting the number of signals for an undamped exponential model using cross-validation approach," *Signal Processing*, vol. 80, no. 3, pp. 525–534, 2000.

[49] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals when the noise covariance matrix is arbitrary," *Journal of Multivariate Analysis*, vol. 20, no. 1, pp. 26–49, 1986.

[50] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1652–1663, Apr. 1993.

[51] J. J. Fuchs, "Estimation of the number of signals in the presence of unknown correlated sensor noise," *IEEE Trans. Signal Process.*, vol. 40, no. 5, pp. 1053–1061, May 1992.

[52] W. B. Bishop and P. M. Djuric, "Model order selection of damped sinusoids in noise by predictive densities," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 611–619, Mar. 1996.

[53] P. Stoica, H. Li, and J. li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, 2000.

[54] C. R. Rao and L. C. Zhao, "Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals," *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 1461–1464, Mar. 1993.

[55] P. Stoica and A. Nehorai, "Study of the statistical performance of the Pisarenko harmonic decomposition method," *IEE Proceedings Radar and Signal Processing*, vol. 135, no. 2, pp. 161–168, Apr. 1988.

[56] P. Stoica and T. Söderström, "Statistical Analysis of MUSIC and Subspace Rotation Estimates of Sinusoidal Frequencies," *IEEE Trans. Signal Process.*, vol. 39, pp. 1836–1847, Aug. 1991.

**Roland Badeau and François Roueff**
roland.badeau@telecom-paris.fr

[57] A. Kot, S. Parthasarathy, D. Tufts, and R. Vaccaro, "The statistical performance of state-variable balancing and Prony's method in parameter estimation," in *Proc. of ICASSP'87*, Apr. 1987, vol. 12, pp. 1549–1552.

[58] A. Eriksson, P. Stoica, and T. Soderstrom, "Second-order properties of MUSIC and ESPRIT estimates of sinusoidal frequencies in high SNR scenarios," *IEE Proceedings on Radar, Sonar and Navigation*, vol. 140, no. 4, pp. 266–272, Aug. 1993.

[59] Y. Hua and T. K. Sarkar, "On SVD for estimating generalized eigenvalues of singular matrix pencil in noise," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 892–900, Apr. 1991.

# Tutorial on non-parametric estimation
# TSIA202b

Roland Badeau

## Exercise 1 *(White noise & Periodogram)*

Let $y_t$ be a (centered) Gaussian white noise of variance $\sigma^2$ and let

$$Y(\nu_k) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} y_t e^{-2i\pi\nu_k t}; \quad \nu_k = \frac{k}{N} \quad k \in \{0,\dots,N-1\}$$

where $Y(\nu_k)$ represents the discrete Fourier transform (DFT) normalized (and evaluated at the Fourier frequency $\nu_k$).

1. Compute the covariances
$$\mathbb{E}\left(Y(\nu_k)\overline{Y(\nu_r)}\right), k,r = 0,\dots,N-1$$
2. Use the result of the previous question to conclude that the periodogram $\hat{S}_{P,yy}(\nu_k) = |Y(\nu_k)|^2$ is an unbiased estimator of the power spectral density (PSD) of $y_t$.
3. Explain why this estimator remains unbiased, even in the case where $\nu \neq \nu_k$ (give an intuitive explanation).

## Exercise 2 *(Reduction of the mean squared error (MSE))*

Let $\hat{x}$ be an estimate of an unknown parameter $x$. Suppose that $\hat{x}$ is unbiased (i.e. $\mathbb{E}(\hat{x}) = x$) and let $\sigma^2_{\hat{x}}$ be the mean square error (MSE) of $\hat{x}$:

$$MSE(\hat{x}) := \sigma^2_{\hat{x}} := \mathbb{E}\left((\hat{x}-x)^2\right)$$

(it should be noted that since $\hat{x}$ is unbiased, $\sigma^2_{\hat{x}}$ is equal to the variance of $\hat{x}$). For a fixed constant $\rho$, we define:

$$\tilde{x} = \rho\hat{x}$$

another estimator of $x$. The coefficient $\rho$, often called the "reduction or regression coefficient", can be chosen in such a way as to minimize the MSE of $\tilde{x}$ (denoted $\sigma^2_{\tilde{x}}$) and to make the latter smaller than $\sigma^2_{\hat{x}}$.

1. *(General result)* Let $\hat{y}$ be an estimator (biased or not) of an unknown parameter $y$. Prove that:

$$MSE(\hat{y}) = \mathbb{E}\left((\hat{y} - \mathbb{E}(\hat{y}))^2\right) + (\mathbb{E}(\hat{y}) - y)^2$$

2. Deduce that the MSE of $\tilde{x}$ reaches its minimum value:

$$\sigma_{\tilde{x}_0}^2 = \rho_0 \sigma_{\hat{x}}^2$$

for:

$$\rho_0 = \frac{x^2}{x^2 + \sigma_{\hat{x}}^2}$$

We will now apply this result to the case of the periodogram. In the course, we have seen that the spectral estimation based on the periodogram $\hat{S}_{P,XX}$ is asymptotically unbiased and has an MSE equal to the square of the value of the power spectral density (PSD) $S_{XX}$:

$$\mathbb{E}\left(\hat{S}_{P,XX}(\nu)\right) \underset{N\to\infty}{\longrightarrow} S_{XX}(\nu), \quad \mathbb{E}\left(\left(\hat{S}_{P,XX}(\nu) - S_{XX}(\nu)\right)^2\right) \underset{N\to\infty}{\longrightarrow} S_{XX}(\nu)^2$$

3. Show that the estimate of the "optimal periodogram" (in the sense mentioned above) is:

$$\tilde{S}_{P,XX}(\nu) = \hat{S}_{P,XX}(\nu)/2$$

and that the MSE of $\tilde{S}_{P,XX}(\nu)$ is equal to half the MSE of $\hat{S}_{P,XX}(\nu)$. Then propose a method to estimate the PSD of the signal from this new estimator minimizing the MSE.

## Exercise 3 *(Estimation of the asymptotic maximum likelihood of $\hat{S}_{P,XX}(\nu)$ from $\hat{S}_{P,XX}(\nu)$)*

It has been proven in the course that, asymptotically w.r.t. $N$, the average of $\hat{S}_{P,XX}(\nu)$ is $S_{XX}(\nu)$ and its variance is $S_{XX}(\nu)^2$. In this exercise, we assume that the $\hat{S}_{P,XX}(\nu)$ are Gaussian. This results in the fact that the probability density of $\hat{S}_{P,XX}(\nu)$ is (for the sake of clarity, we omit the dependency in $\nu$):

$$p_{S_{XX}}\left(\hat{S}_{P,XX}\right) = \frac{1}{\sqrt{2\pi S_{XX}^2}} \exp\left(-\frac{\left(\hat{S}_{P,XX} - S_{XX}\right)^2}{2S_{XX}^2}\right)$$

1. Show that the maximum likelihood estimate (MLE) of $S_{XX}$ by $\hat{S}_{P,XX}$ (defined as the maximum argument $S_{XX}$ of $p_{S_{XX}}\left(\hat{S}_{P,XX}\right)$ ) is given by:

$$\tilde{S}_{P,XX} = \frac{-1+\sqrt{5}}{2}\hat{S}_{P,XX}$$

2. Compare $\tilde{S}_{P,XX}$ with the estimate of $S_{XX}$ presented in Exercise 2, in terms of bias and MSE.

# Tutorial on parametric estimation of rational spectra
# TSIA202b

Roland Badeau

## Exercise 1: Maximum entropy distribution

Let $X$ be a real random variable. Prove that the probability distribution $p_X$ that maximizes the entropy of $X$, subject to fixed mean $\mu_X$ and variance $\sigma_X^2$, is the Gaussian distribution.

## Exercise 2: Relationship between AR Modeling and Forward Linear Prediction

Suppose we have a zero-mean WSS process $\{Y_t\}$ (not necessarily AR) with autocovariance function $\{r_{YY}(k)\}_{k=-\infty}^{\infty}$. We wish to predict $Y_t$ by a linear combination of its $p$ past values: the predicted value is given by

$$\hat{Y}_t^f = \sum_{k=1}^{p} a_k Y_{t-k}$$

We define the forward prediction error as

$$Z_t^f = Y_t - \hat{Y}_t^f = Y_t - \sum_{k=1}^{p} a_k Y_{t-k}.$$

Show that the vector $\theta_f = [a_1 \dots a_p]^\top$ of prediction coefficients that minimizes the prediction-error variance $\sigma_f^2 \triangleq \mathbb{E}\{|Z_t^f|^2\}$ is the solution to

$$\begin{bmatrix} r_{YY}(0) & r_{YY}(-1) & \dots & r_{YY}(-p) \\ r_{YY}(1) & r_{YY}(0) & & \vdots \\ \vdots & & \ddots & r_{YY}(-1) \\ r_{YY}(p) & \dots & & r_{YY}(0) \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} \sigma_p^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1}$$

Show also that $\sigma_f^2 = \sigma_p^2$ (i.e., that $\sigma_p^2$ in (1) is the prediction-error variance).

## Exercise 3: Relationship between AR Modeling and Backward Linear Prediction

Consider the signal $\{Y_t\}$, as in Exercise 2. This time, we will consider backward prediction: we will predict $Y_t$ from its $p$ immediate future values:

$$\hat{Y}_t^b = \sum_{k=1}^{p} b_k Y_{t+k}$$

with corresponding backward prediction error $Z_t^b = Y_t - \hat{Y}_t^b$. Such backward prediction is useful in applications where noncausal processing is permitted; for example, when the data has been prerecorded and is stored in memory or on a tape and we want to make inferences on samples that precede the observed ones. Find an expression similar to (1) for the backward prediction coefficient vector $\theta_b = [b_1 \dots b_p]^\top$. Find a relationship between the $\theta_b$ and the corresponding forward prediction coefficient vector $\theta_f$. Relate the forward and backward prediction error variances.

## Exercise 4: Prediction Filters and Smoothing Filters

The smoothing filter is a practically useful variation on the theme of linear prediction. A result of Exercises 2 and 3 should be that for the forward and backward prediction filters

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \text{ and } B(z) = 1 + \sum_{k=1}^{p} b_k z^{-k},$$

the prediction coefficients satisfy $a_k = \overline{b_k}$, and the prediction error variances are equal. Now consider the smoothing filter

$$Z_t^s = Y_t - \sum_{k=1}^{p} c_k Y_{t-k} - \sum_{k=1}^{p} d_k Y_{t+k}.$$

1. Derive a system of linear equations, similar to the forward and backward linear prediction equations, that relate the smoothing filter coefficients, the smoothing prediction error variance $\sigma_s^2 = \mathbb{E}\left\{|Z_t^s|^2\right\}$, and the autocovariance function of $Y_t$.

2. Show that the unconstrained minimum smoothing error variance solution satisfies $c_k = \overline{d_k}$.

3. Prove that the minimum smoothing prediction error variance is less than the minimum (forward or backward) prediction error variance.

## Exercise 5: Generating the autocovariance function from ARMA parameters

In this lesson we have expressed the ARMA coefficients $\sigma_Z^2, a_i, b_j$ in terms of the autocovariance function $\{r_{XX}(k)\}_{k=-\infty}^{\infty}$. Find the inverse map: given $\sigma_Z^2, a_1, \dots, a_p, b_1 \dots, b_q$, find equations to determine $\{r_{XX}(k)\}_{k=-\infty}^{\infty}$.

# Tutorial on filter bank methods
# TSIA202b

Roland Badeau

## Exercise 1: Capon Estimate of the Parameters of a Single Sine Wave

Assume that the data under study consists of a sinusoidal signal observed in white noise: $\forall t \in \mathbb{Z}$,

$$X_t = \alpha e^{i(2\pi\nu_0 t+\varphi)} + Z_t$$

where $\alpha \in \mathbb{C}$, $\nu_0 \in \mathbb{R}$, $\varphi$ is a random variable uniformly distributed on $[0, 2\pi]$, and $Z_t \sim BB(0, \sigma^2)$ is independent from $\varphi$.

1. Prove that $X_t$ is a WSS process and give the mathematical expression of its mean $\mu_X$ and its autocovariance function $r_{XX}(k)$.
2. Explain the expression *line spectrum* that is used to name the spectral measure of $X_t$.
3. Show that the $N \times N$ covariance matrix $R_{XX}$ of $X_1 \dots X_N$ is given by:

   $$R_{XX} = |\alpha|^2 e(\nu_0) e(\nu_0)^H + \sigma^2 I_N$$

   where $e(\xi) = [1, e^{i2\pi\xi} \dots e^{i2\pi\xi(N-1)}]^\top$ and $I_N$ denotes the $N \times N$ identity matrix.
4. Check that $\left\{ \frac{e(\nu_0 + \frac{k}{N})}{\sqrt{N}} \right\}_{k \in \{0 \dots N-1\}}$ forms a unitary basis of eigenvectors of $R_{XX}$ and give the expressions of the corresponding eigenvalues.
5. Prove that the Capon spectrum $\hat{S}_{\mathrm{CAP},XX}(\nu) = \frac{N}{e(\nu)^H R_{XX}^{-1} e(\nu)}$ peaks at $\nu = \nu_0$, and show that the height of the peak is $N|\alpha|^2 + \sigma^2$.

Institut Mines-Télécom

# Exercises on high resolution methods

**Roland Badeau**
roland.badeau@telecom-paristech.fr

Let us consider the *Exponential Sinusoidal Model* (ESM) :

$$s[t] = \sum_{k=0}^{K-1} a_k \, e^{\delta_k t} e^{i(2\pi f_k t + \phi_k)},$$

which, to each frequency $f_k \in \left]-\frac{1}{2}, \frac{1}{2}\right]$, associates a real amplitude $a_k > 0$, a phase $\phi_k \in ]-\pi, \pi]$, and a damping factor $\delta_k \in \mathbb{R}$. By defining the complex amplitudes $\alpha_k = a_k e^{i\phi_k}$ and the complex poles $z_k = e^{\delta_k + i2\pi f_k}$, this model can be rewritten in the more compact form

$$s[t] = \sum_{k=0}^{K-1} \alpha_k \, {z_k}^t.$$

In practice, the observed signal $x[t]$ never exactly fits this model. It is rather modeled as the sum of signal $s[t]$ plus a complex Gaussian white noise $b[t]$ of variance $\sigma^2$ :

$$x[t] = s[t] + b[t].$$

**Remark :** A complex Gaussian white noise of variance $\sigma^2$ is a complex process whose real part and imaginary part are two Gaussian white noises of same variance $\frac{\sigma^2}{2}$, independent from each other.

We assume that the signal $x[t]$ is observed on the time interval $\{0 \dots N-1\}$ of length $N > 2K$. We then consider two integers $n$ and $l$ such that $n > K$, $l > K$, and $N = n + l - 1$.

Finally, we define the $n \times l$ Hankel matrix which contains the $N$ samples of the observed signal :

$$\boldsymbol{X} = \begin{bmatrix} x[0] & x[1] & \dots & x[l-1] \\ x[1] & x[2] & \dots & x[l] \\ \vdots & \vdots & \vdots & \vdots \\ x[n-1] & x[n] & \dots & x[N-1] \end{bmatrix}.$$

We define in the same way the Hankel matrices $\boldsymbol{S}$ and $\boldsymbol{B}$ of same dimension $n \times l$, from the samples of $s[t]$ and $b[t]$, respectively.

**Notation :**

— $\boldsymbol{X}^T$ : transpose of matrix $\boldsymbol{X}$,
— $\boldsymbol{X}^*$ : conjugate of matrix $\boldsymbol{X}$,
— $\boldsymbol{X}^H$ : Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{X}$.

# 1 *Mu*ltiple *Si*gnal *C*lassification (MUSIC)

**Question 1** For all $k \in \{0 \dots K-1\}$, we consider the component $s_k[t] = \alpha_k \, {z_k}^t$. We then define the $n \times l$ Hankel matrix

$$\boldsymbol{S_k} = \begin{bmatrix} s_k[0] & s_k[1] & \dots & s_k[l-1] \\ s_k[1] & s_k[2] & \dots & s_k[l] \\ \vdots & \vdots & \vdots & \vdots \\ s_k[n-1] & s_k[n] & \dots & s_k[N-1] \end{bmatrix}$$

For all $z \in \mathbb{C}$, let us define the $n$-dimensional vector $\boldsymbol{v}^n(z) = [1, z, z^2, \dots, z^{n-1}]^\top$, and the $l$-dimensional vector $\boldsymbol{v}^l(z) = [1, z, z^2, \dots, z^{l-1}]^\top$. Then prove that $\boldsymbol{S_k} = \alpha_k \, \boldsymbol{v}^n(z_k) \, \boldsymbol{v}^l(z_k)^\top$.

**Roland Badeau** roland.badeau@telecom-paristech.fr

Contexte académique } sans modifications
*Voir page 3*

**Question 2** Use the result of question 1 to prove that $\boldsymbol{S} = \sum_{k=0}^{K-1} \alpha_k \, \boldsymbol{v}^n(z_k) \, \boldsymbol{v}^l(z_k)^\top$. Prove that this last equality can be rewritten in the form $\boldsymbol{S} = \boldsymbol{V}^n \, \boldsymbol{A} \, {\boldsymbol{V}^l}^\top$, where

— $\boldsymbol{V}^n$ is an $n \times K$ Vandermonde matrix :

$$\boldsymbol{V}^n = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ {z_0}^2 & {z_1}^2 & \dots & {z_{K-1}}^2 \\ \vdots & \vdots & \vdots & \vdots \\ {z_0}^{n-1} & {z_1}^{n-1} & \dots & {z_{K-1}}^{n-1} \end{bmatrix}$$

— $\boldsymbol{V}^l$ is an $l \times K$ Vandermonde matrix,

— $\boldsymbol{A} = \mathrm{diag}(\alpha_0, \alpha_1, \dots, \alpha_{K-1})$ is a $K \times K$ diagonal matrix.

**Question 3** Let us define matrix $\boldsymbol{R}_{ss} = \frac{1}{l} \boldsymbol{S} \boldsymbol{S}^H$. Prove that matrix $\boldsymbol{R}_{ss}$ is Hermitian and positive semi-definite. Prove that $\boldsymbol{R}_{ss}$ can be factorized in the form $\boldsymbol{R}_{ss} = \boldsymbol{V}^n \, \boldsymbol{P} \, {\boldsymbol{V}^n}^H$, where $\boldsymbol{P}$ is a $K \times K$ Hermitian and positive definite matrix. Conclude that matrix $\boldsymbol{R}_{ss}$ has rank $K$ (we remind that the poles $z_k$ are pairwise distinct).

**Question 4** Prove that matrix $\boldsymbol{R}_{ss}$ is diagonalizable in an orthonormal basis, and that its eigenvalues $\{\lambda_i\}_{i=0\dots n-1}$ are non-negative. By assuming that they are sorted in decreasing order and by using the result of question 3, conclude that

— $\forall i \in \{0 \dots K-1\}, \lambda_i > 0$ ;

— $\forall i \in \{K \dots n-1\}, \lambda_i = 0$.

**Question 5** Let $\widehat{\boldsymbol{R}}_{xx} = \frac{1}{l} \boldsymbol{X} \boldsymbol{X}^H$ and $\boldsymbol{R}_{xx} = \mathbb{E}\left[\widehat{\boldsymbol{R}}_{xx}\right]$. Similarly, let $\widehat{\boldsymbol{R}}_{bb} = \frac{1}{l} \boldsymbol{B} \boldsymbol{B}^H$ and $\boldsymbol{R}_{bb} = \mathbb{E}\left[\widehat{\boldsymbol{R}}_{bb}\right]$. By using equality $\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{B}$ and the fact that the noise is centered, prove that $\boldsymbol{R}_{xx} = \boldsymbol{R}_{ss} + \boldsymbol{R}_{bb}$. Then prove that for a complex Gaussian white noise, $\boldsymbol{R}_{bb} = \sigma^2 \boldsymbol{I}_n$.

**Question 6** For all $i \in \{0 \dots n-1\}$, let $\boldsymbol{w}_i$ denote the eigenvector of matrix $\boldsymbol{R}_{ss}$ corresponding to the eigenvalue $\lambda_i$. By using the result of question 5, prove that $\boldsymbol{w}_i$ is also an eigenvector of $\boldsymbol{R}_{xx}$ corresponding to the eigenvalue $\lambda'_i = \lambda_i + \sigma^2$. Conclude that

— $\forall i \in \{0 \dots K-1\}, \lambda'_i > \sigma^2$ ;

— $\forall i \in \{K \dots n-1\}, \lambda'_i = \sigma^2$.

**Question 7** Let $\boldsymbol{W}$ denote the matrix $[\boldsymbol{w}_0 \dots \boldsymbol{w}_{K-1}]$, and $\boldsymbol{W}_\perp$ the matrix $[\boldsymbol{w}_K \dots \boldsymbol{w}_{n-1}]$. Prove that $\mathrm{Span}(\boldsymbol{W}) = \mathrm{Span}(\boldsymbol{V}^n)$ (you can start by proving that $\mathrm{Span}(\boldsymbol{W}) \subset \mathrm{Span}(\boldsymbol{V}^n)$).

**Remark :** The subspace spanned by $\boldsymbol{W}_\perp$ is an eigen-subspace of matrix $\boldsymbol{R}_{xx}$ corresponding to the eigenvalue $\sigma^2$. This is why it is called *noise subspace*. The orthonormal matrix $\boldsymbol{W}$ and the Vandermonde matrix $\boldsymbol{V}^n$ span the same subspace. It thus completely characterizes the $K$ poles of the signal, This is why it is called *signal subspace*. However, all the eigenvalues of $\boldsymbol{R}_{xx}$ corresponding to the signal subspace are increased by $\sigma^2$, which means that this subspace also contains noise.

**Question 8** Prove that the poles $\{z_k\}_{k\in\{0\dots K-1\}}$ are the solutions of equation $\left\|\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)\right\|^2 = 0$.

**Roland Badeau** roland.badeau@telecom-paristech.fr



*Voir page 3*

**Remark :** In practice, real signals do not rigorously fit the model, and this equation does never hold exactly. This is why the "spectral-MUSIC" method for estimating the poles consists in detecting the $K$ highest peaks of function $z \mapsto \frac{1}{\|\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)\|^2}$. It is thus easier to implement than the maximum likelihood method, which requires the numerical optimization of a cost function of $K$ complex variables.

## 2 *E*stimation of *S*ignal *P*arameters via *R*otational *I*nvariance *T*echniques (ESPRIT)

Let $\boldsymbol{V}_\downarrow^n$ be the $(n-1)\times K$ matrix that contains the $n-1$ first rows of $\boldsymbol{V}^n$, and $\boldsymbol{V}_\uparrow^n$ the $(n-1)\times K$ matrix that contains the $n-1$ last rows of $\boldsymbol{V}^n$. Similarly, let $\boldsymbol{W}_\downarrow$ be the $(n-1)\times K$ matrix that contains the $n-1$ first rows of $\boldsymbol{W}$, and $\boldsymbol{W}_\uparrow$ the $(n-1)\times K$ matrix that contains the $n-1$ last rows of $\boldsymbol{W}$.

**Question 1** Prove that matrices $\boldsymbol{V}_\downarrow^n$ and $\boldsymbol{V}_\uparrow^n$ are such that $\boldsymbol{V}_\uparrow^n = \boldsymbol{V}_\downarrow^n \boldsymbol{D}$, where $\boldsymbol{D}$ is a $K \times K$ diagonal matrix. What are its diagonal entries ?

**Question 2** Prove that there is a $K \times K$ invertible matrix $\boldsymbol{G}$ such that $\boldsymbol{V}^n = \boldsymbol{W}\boldsymbol{G}$ (we do not ask to compute $\boldsymbol{G}$, but only to prove its existence). Then prove that $\boldsymbol{V}_\downarrow^n = \boldsymbol{W}_\downarrow \boldsymbol{G}$ and $\boldsymbol{V}_\uparrow^n = \boldsymbol{W}_\uparrow \boldsymbol{G}$.

**Question 3** Conclude that there is an invertible matrix $\boldsymbol{\Phi}$ such that $\boldsymbol{W}_\uparrow = \boldsymbol{W}_\downarrow \boldsymbol{\Phi}$. What are the eigenvalues of $\boldsymbol{\Phi}$ ?

**Question 4** By assuming that matrix $\boldsymbol{W}_\downarrow^H \boldsymbol{W}_\downarrow$ is invertible, compute $\boldsymbol{\Phi}$ as a function of $\boldsymbol{W}_\downarrow$ and $\boldsymbol{W}_\uparrow$.

**Question 5** Propose an estimation method of the poles $\{z_k\}_{k\in\{0\ldots K-1\}}$.

Institut Mines-Télécom

# Practical work on spectral estimation

**Roland Badeau**

In this practical work, we propose to compare nonparametric spectral estimators (periodogram, smoothed periodogram) with parametric estimators (linear prediction, Capon, Durbin methods), in the case of three types of processes commonly encountered : autoregressive processes, sum of noisy sinusoids, and autoregressive-moving-average processes.

This practical work will be carried out either with Matlab or Python. The different estimators will be applied to synthetic signals, as well as speech signals.

# 1 Autoregressive process

## 1.1 Synthesis

1. Code a function that synthesizes an autoregressive (AR) process of length $N$, defined by its AR coefficients and the innovation variance $\sigma^2$.

## 1.2 Non-parametric estimation : periodogram

2. Code a function that computes and displays the periodogram of a given input signal, allowing to choose between the standard method and the Bartlett and Welsh methods.
3. Display the periodograms obtained with the different methods (in dB), in the case of an AR process of order 1 whose parameters can be chosen by hand. We will superimpose in each case the periodogram with the known power spectral density (PSD) of the AR process. We will pay attention to the correct normalization of the different estimators.
4. Experimentally compare these different estimators in terms of bias and variance.

## 1.3 Parametric estimation : linear prediction

5. Estimate the AR model using either the code you wrote in the previous practical work of TSIA202a, or the function `lpc` (linear prediction method) from Matlab or from the librosa Python package.
6. Superimpose the spectral envelope of the estimated AR model with that of the exact AR model and with the periodogram. Comment.

## 1.4 Application to speech signals

7. Apply the above methods to a vowel sound (such as 'a' for example, that we can extract from file `aeiou.wav`, downloadable from the eCampus website of TSIA202b), and superimpose the estimated spectra. Comment.

# 2 Sum of noisy sinusoids

## 2.1 Synthesis

1. Code a function that synthesizes a signal of length $N$, consisting of the sum of $K$ real sinusoids defined by their frequencies, amplitudes and origin phases, and a Gaussian white noise of variance $\sigma^2$.

### 2.2 Non parametric estimation : periodogram

2. Display the periodograms obtained with the standard method and the Bartlett and Welsh methods, in the case of $K = 2$ sinusoidal frequencies $\nu_1$ and $\nu_2$, zero origin phase and same amplitude equal to 1. We will superimpose in each case the periodogram with two vertical lines centered at frequencies $\nu_1$ and $\nu_2$, and a horizontal line centered at the variance $\sigma^2$.
3. Compare the resolution of the various periodograms, by varying the difference $\Delta\nu = |\nu_1 - \nu_2|$ between both frequencies, as well as the signal-to-noise ratio (SNR) in dB.

### 2.3 Parametric estimation : linear prediction

4. Estimate a fourth order AR model using the linear prediction method.
5. Superimpose the estimated AR model with the previous figures, always by varying the gap $\Delta\nu$, as well as the SNR in dB. Comment on the results in terms of resolution.

### 2.4 Parametric estimation : Capon estimator

6. Code a function that implements the Capon estimator.
7. Compare the resolution of this method with the two previous methods, always by varying $\Delta\nu$, as well as the SNR in dB.

## 3 Autoregressive-moving-average process

### 3.1 Synthesis

1. Code a function that synthesizes an autoregressive-moving-average (ARMA) process of length $N$, defined by its AR and MA coefficients and by the innovation variance $\sigma^2$.

### 3.2 Non-parametric estimation : periodogram

2. Display the periodograms obtained with the different methods, in the case of an ARMA process of order $(1, 1)$ whose parameters can be chosen by hand. We will superimpose in each case the periodogram with the known PSD of the ARMA process.
3. Compare these different estimators in terms of bias and variance.

### 3.3 Parametric estimation : Durbin method

4. Code a function that implements the Durbin method (estimation of the AR part by shifting the autocorrelation, estimation of the MA part via a long AR).
5. Superimpose the spectral envelope of the estimated ARMA model with that of the exact ARMA model and with the periodogram. Comment.

### 3.4 Application to speech signals

6. Apply the above methods to a nasal vowel sound (such as 'in' for example, that we will be able to extract from file `an_in_on.wav`, downloadable on the eCampus website of TSIA202b), and superimpose the estimated spectra.

Contexte académique } sans modifications

Institut Mines-Télécom

# Analysis and synthesis of bell sounds

**Roland Badeau**

The various files related to this practical work can be downloaded on the eCampus website of TSIA202b. You can load a sound file with Matlab by using the command `[x,Fs] = wavread('clocheB.wav')`. In order to listen to it, you can type up `soundsc(x,Fs)`. In Python, you can use the provided notebook template `template-TP-HR.ipynb`.

# 1 Introduction

Bells are among the oldest music instruments and the sound they produce is often evocative because it has soothed the everyday life of generations for about 3000 years, accompanying minor and major events. This evocation is partly due to the structure of the sound spectrum: the eigenmodes of vibration are generally tuned by the bell makers so that their frequencies follow a particular series, which includes the minor third (E flat if the bell is tuned in C). This series is not harmonic, but the ratios between the eigenfrequencies are such that one can perceive a well-defined pitch. In particular, the presence of the series 2-3-4, strong at the beginning of the sound, reinforces the feeling of pitch in the neighborhood of the fundamental frequency. This feeling is related to a psychoacoustic effect (processing of the signal received by the brain).

Let $f_p$ be the frequency corresponding to the perceived pitch. The analysis of the eigenfrequencies series leads to a table of about 15 ratios $\alpha_n = f_n/f_p$. Their orders of magnitude are as follows: 0.5 (hum / "*bourdon*"), 1 (prime / fundamental), 1.2 (minor third), 1.5 (fifth), 2 (nominal / octave), 2.5, 2.6, 2.7, 3, 3.3, 3.7, 4.2 (wrong double octave), 4.5, 5, 5.9. The timbre of the corresponding sound depends on the amplitude and on the decrease of each of these partials.

This practical work aims to develop a high resolution spectral estimation method in order to perform the analysis / synthesis of bell sounds. As can be noticed in figure 1, this type of sounds presents a strong temporal decrease.

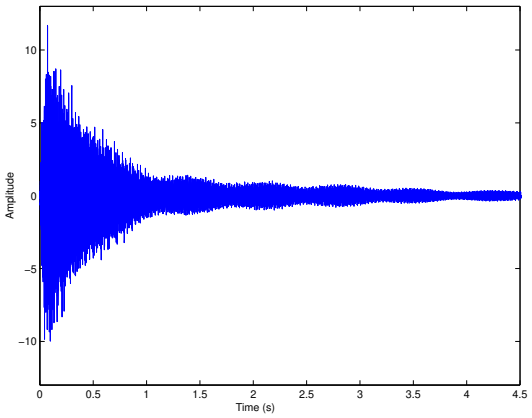



Figure 1: Bell sound

In order to take this damping into account, we consider the *Exponential Sinusoidal Model* (ESM):

$$s[t] = \sum_{k=0}^{K-1} a_k \, e^{\delta_k t} e^{i(2\pi f_k t + \phi_k)},$$

which to each frequency $f_k \in \left]-\frac{1}{2}, \frac{1}{2}\right]$ associates a real amplitude $a_k > 0$, a phase $\phi_k \in ]-\pi, \pi]$, and

a damping factor $\delta_k \in \mathbb{R}$. By defining the complex amplitudes $\alpha_k = a_k e^{i\phi_k}$ and the complex poles $z_k = e^{\delta_k + i2\pi f_k}$, this model can be rewritten in the form

$$s[t] = \sum_{k=0}^{K-1} \alpha_k \, {z_k}^t.$$

The case $\delta_k < 0$ corresponds to exponentially decreasing sinusoids, which are solutions of physical propagation equations. The model parameters are then $\{\delta_k, f_k, a_k, \phi_k\}_{k\in\{0\ldots K-1\}}$. In order to estimate them, we will use the ESPRIT method presented in the course. Firstly, we will apply it to a synthetic signal in order to highlight the superiority of high resolution methods over Fourier analysis in terms of spectral resolution. Then this method will be applied to bell sounds.

# 2 Reminder about Matlab/Python

In Matlab:

- `A'`: Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{A}$;
- `A.'`: transpose of matrix $\boldsymbol{A}$ (without complex conjugation);
- `A(l1:l2,c1:c2)`: matrix extracted from $\boldsymbol{A}$ between rows $l_1$ and $l_2$ (inclusive) and columns $c_1$ and $c_2$ (inclusive).

In Python:

- `A.conj().T`: Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{A}$;
- `A.T`: transpose of matrix $\boldsymbol{A}$ (without complex conjugation);
- `A[l1:l2,c1:c2]`: matrix extracted from $\boldsymbol{A}$ between rows $l_1$ and $l_2$ (excluded) and columns $c_1$ and $c_2$ (excluded).

# 3 Synthetic signal

Here we consider a synthetic signal of length $N$, consisting of a sum of two complex exponentials, whose frequencies are separated by an interval $\Delta f = \frac{1}{N}$ (which corresponds to the resolution limit of Fourier analysis). The phases are drawn randomly, according to a uniform probability distribution on $(-\pi, \pi)$. We do not add noise to this signal, so that that the observed signal $x[t]$ is equal to $s[t]$. You can use the following parameters: $N = 63$, $f_0 = \frac{1}{4}$, $f_1 = f_0 + \frac{1}{N}$, $a_0 = 1$, $a_1 = 10$, $\delta_0 = 0$, $\delta_1 = -0.05$. In order to synthesize it, you can call the provided function

```
x = Synthesis(N,delta,f,a,phi);
```

whose arguments are $N$, the vector `delta` of damping factors $\delta_k$, the vector `f` of frequencies $f_k$, the vector `a` of amplitudes $a_k$, and the vector `phi` of phases $\phi_k$.

## 3.1 Spectral analysis by Fourier transform

Observe the periodogram of this signal. Briefly study the separability of the two spectral lines, without zero-padding ($N_{\text{fft}} = N$) and with zero-padding ($N_{\text{fft}} = 1024 > N$).

## 3.2 High resolution methods

Our goal is to write functions

`MUSIC(x,n,K)` and `[delta,f] = ESPRIT(x,n,K)`

which analyze the signal $x$ of length $N$ by using methods MUSIC and ESPRIT, with a signal subspace of dimension $K$ and a noise subspace of dimension $n-K$, and the data vectors of length $n$ ranging from $K+1$ to $N-K+1$. In order to process the synthetic signals, you can choose $n = 32$ and $K = 2$. The two methods share the following steps:

1. **Computation of the empirical covariance matrix**

   The empirical covariance matrix of the observed signal is defined by the equation

   $$\widehat{\boldsymbol{R}}_{xx} = \frac{1}{l}\boldsymbol{X}\,\boldsymbol{X}^H$$

   where $\boldsymbol{X}$ is an $n \times l$ Hankel matrix containing the $N = n + l - 1$ samples of the signal:

   $$\boldsymbol{X} = \begin{bmatrix} x[0] & x[1] & \dots & x[l-1] \\ x[1] & x[2] & \dots & x[l] \\ \vdots & \vdots & \ddots & \vdots \\ x[n-1] & x[n] & \dots & x[N-1] \end{bmatrix}$$

   Matrix $\boldsymbol{X}$ can be constructed with function `hankel`.

2. **Estimation of the signal subspace**

   Matrix $\widehat{\boldsymbol{R}}_{xx}$ can be diagonalized with the command `[U1,Lambda,U2] = svd(Rxx)`. Matrix $\widehat{\boldsymbol{R}}_{xx}$ being positive semidefinite, the column vectors of the $n \times n$ matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are the eigenvectors of $\widehat{\boldsymbol{R}}_{xx}$, corresponding to the $n$ eigenvalues sorted in the diagonal matrix $\boldsymbol{\Lambda}$ in decreasing order (we thus have $\widehat{\boldsymbol{R}}_{xx} = \boldsymbol{U}_1\,\boldsymbol{\Lambda}\,\boldsymbol{U}_1^H = \boldsymbol{U}_2\,\boldsymbol{\Lambda}\,\boldsymbol{U}_2^H$). Therefore you can extract from $\boldsymbol{U}_1$ (or from $\boldsymbol{U}_2$) an $n \times K$ basis of the signal subspace $\boldsymbol{W}$.

### 3.2.1 ESPRIT method

In a first stage, the ESPRIT method consists in estimating the frequencies and damping factors:

3. **Estimation of the frequencies and damping factors**

   In order to estimate the frequencies, you can proceed in the following way:

   - extract from $\boldsymbol{W}$ the matrices $\boldsymbol{W}_\downarrow$ (obtained by removing the last row of $\boldsymbol{W}$) and $\boldsymbol{W}_\uparrow$ (obtained by removing the first row of $\boldsymbol{W}$);
   - compute $\boldsymbol{\Phi} = \left(\left(\boldsymbol{W}_\downarrow^H \boldsymbol{W}_\downarrow\right)^{-1} \boldsymbol{W}_\downarrow^H\right) \boldsymbol{W}_\uparrow = \boldsymbol{W}_\downarrow^\dagger \boldsymbol{W}_\uparrow$, where the symbol † denotes the pseudo inverse operator (Matlab function `pinv` or Python function `numpy.linalg.pinv`).
   - compute the eigenvalues of $\boldsymbol{\Phi}$ by using the Matlab function `eig` or the Python function `numpy.linalg.eig` (we remind that the eigenvalues of $\boldsymbol{\Phi}$ are the poles $z_k = e^{\delta_k + i2\pi f_k}$). Then compute $\delta_k = \ln(|z_k|)$ and $f_k = \frac{1}{2\pi}\,\mathrm{angle}(z_k)$.

4. **Estimation of the amplitudes and phases**

   We now aim to write a function

   ```
   [a,phi] = LeastSquares(x,delta,f)
   ```

   which estimates the amplitudes $a_k$ and phases $\phi_k$ by means of the least squares method, given the signal $x$, the damping factors $\delta_k$ and the frequencies $f_k$. The complex amplitudes are thus determined by the equation

   $$\boldsymbol{\alpha} = \left( (\boldsymbol{V}^{N\,H} \boldsymbol{V}^N)^{-1} \boldsymbol{V}^{N\,H} \right) \boldsymbol{x} = \boldsymbol{V}^{N\,\dagger}\, \boldsymbol{x} \tag{1}$$

   where $\boldsymbol{x}$ is the vector $[x[0], \ldots, x[N-1]]^\top$ and $\boldsymbol{V}^N$ is the $N \times K$ Vandermonde matrix, whose entries are such that $\boldsymbol{V}^N_{(t,k)} = {z_k}^t$ for all $(t,k) \in \{0 \ldots N-1\} \times \{0 \ldots K-1\}$. In order to compute matrix $\boldsymbol{V}^N$, it is possible of avoid using a `for` loop by noting that $\ln\left(\boldsymbol{V}^N_{(t,k)}\right) = t(\delta_k + i2\pi f_k)$. Therefore, the matrix containing the coefficients $\ln\left(\boldsymbol{V}^N_{(t,k)}\right)$ can be expressed as the product of a column vector and a row vector. Then compute $a_k = |\alpha_k|$ and $\phi_k = \text{angle}(\alpha_k)$ for all $k \in \{0 \ldots K-1\}$.

5. **Application to synthetic signals**

   Apply functions `ESPRIT` and `LeastSquares` to the previously synthesized signal. Comment.

### 3.2.2 MUSIC method

We remind that the MUSIC pseudo-spectrum is defined as $P(z) = \dfrac{1}{\|\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)\|^2}$, where matrix $\boldsymbol{W}_\perp$ spans the noise subspace.

6. **MUSIC pseudo-spectrum**

   Write a function `MUSIC(x,n,K)` which plots the logarithm of the pseudo-spectrum as a function of the two variables $f \in [0,1]$ and $\delta \in [-0.1, 0.1]$ (you can use the Matlab function `surf` or the Python Matplotlib function `plot_surface`). Apply function `MUSIC` to the previously synthesized signal, and check that the pseudo-spectrum makes the two poles $z_k = e^{\delta_k + i2\pi f_k}$ clearly visible.

# 4 Audio signals

We now propose to apply the functions developed in the previous part to bell sounds.

## 4.1 Spectral analysis by Fourier transform

Look at the periodogram of the signal `ClocheA.wav` or `ClocheB.wav`, and compare the series of its eigenfrequencies with the values given in the introduction.

## 4.2 High resolution method

We now want to apply the ESPRIT method to this signal. Let $K = 54$, $n = 512$ and $l = 2n = 1024$ (hence $N = n + l - 1 = 1535$).

In order to guarantee that the signal model holds on the analysis window (exponential damping), we will extract a segment of length $N$ whose beginning is posterior to the maximum of the waveform envelope. We may thus start at the $10000^{\text{th}}$ sample.

Apply function `ESPRIT` to the extracted signal in order to estimate the eigenfrequencies and the corresponding damping factors. Then estimate the amplitudes and phases by calling function `LeastSquares`. Finally, listen to the signal resynthesized with function `Synthesis` (on a duration longer than the extracted segment, in order to clearly highlight the sound resonances), and comment.

# Examination of the teaching unit
# *Linear time series* (part 2) - TSIA202b
April 12th 2023

*Duration: 1:30*
*All documents are permitted. However electronic devices (including calculators) are forbidden.*

## Exercise 1: Spectral estimation

We have estimated the first terms of the autocovariance function of a real WSS random process, and we obtained the values $r_{XX}(0) = 1$ and $r_{XX}(1) = \rho$ with $|\rho| < 1$. We now want to estimate its power spectral density (PSD). For each of the following spectral estimators, give its general expression, and its simplified expression in the particular case of this WSS process $X$:

1. periodogram method;
2. maximum entropy method;
3. Capon's method (*minimum variance distortionless response*).

## Exercise 2: Autoregressive modeling

We measure the first samples of a WSS process $X_t$ for $t \in [0 \dots 7]$, and their values are $[-1, 0, 1, 1, 0, -1, -1, 0]$.

1. What is the minimal number of sinusoids allowing us to represent the first 8 samples of this signal? (Hint: you can try to draw a smooth curve which interpolates the samples to help you guess the solution) Give their frequencies, amplitudes and phases.
2. We want to estimate the PSD of this signal by the maximum entropy method. What autoregressive (AR) order $p$ should be selected? (Hint: when a signal is formed of a sum of $N$ real sinusoids, the covariance matrix has rank $2N$, so the AR order $p$ should be $2N$)
3. Give the general expression of the biased estimator of the autocovariance $\hat{r}_{XX}(k)$ (which is guaranteed to be positive definite), and its values in this particular case for $k \in [0, p]$.
4. Give the general expression of Yule-Walker equations. Solve these equations in this particular case and deduce the coefficients of the autoregressive model.
5. Without explicitly calculating the power spectral density of this model, explain how to estimate the frequencies of the peaks of the PSD of any autoregressive process.
6. Apply this method to the model that you have calculated at question 4, and compare the result to that of question 1.

## Exercise 3: High resolution estimation of multiple complex poles

In the lecture on high resolution methods, we considered $K$ distinct and non-zero complex poles $z_0, \ldots, z_{K-1} \in \mathbb{C}$ and we defined the polynomial of degree $K$ whose dominant coefficient is 1 and whose roots are the $z_k$: $P[z] = \prod_{k=0}^{K-1}(z - z_k) = \sum_{\tau=0}^{K} p_\tau \, z^{K-\tau}$. We then proved that a complex discrete time signal $\{s(t)\}_{t\in\mathbb{Z}}$ satisfies the recursive equation $\forall t \in \mathbb{Z}$, $\sum_{\tau=0}^{K} p_\tau \, s(t-\tau) = 0$ if and only if there are $K$ scalars $\alpha_0, \ldots, \alpha_{K-1} \in \mathbb{C}$ such that $s(t) = \sum_{k=0}^{K-1} \alpha_k \, {z_k}^t$ (ESM model). Thus the estimation of the poles $z_k$ can be carried out by linear prediction (Prony and Pisarenko methods). The aim of this exercise is to study the more general case where $P$ has multiple poles:

$$P[z] = \prod_{k=0}^{K-1} (z - z_k)^{M_k} = \sum_{\tau=0}^{M} p_\tau \, z^{M-\tau}, \tag{1}$$

where $\forall k \in \{0 \ldots K-1\}$, $z_k \neq 0$, $M_k \in \mathbb{N}^*$, and $M = \sum_{k=0}^{K-1} M_k$. We will show that a discrete time signal $\{s(t)\}_{t\in\mathbb{Z}}$ satisfies the recurrence equation

$$\forall t \in \mathbb{Z}, \; \sum_{\tau=0}^{M} p_\tau \, s(t-\tau) = 0 \tag{2}$$

if and only if $\forall k \in \{0 \ldots K-1\}$, there is a polynomial $\alpha_k[t] \in \mathbb{C}[t]$ of degree $M_k - 1$ such that

$$s(t) = \sum_{k=0}^{K-1} \alpha_k[t] \, {z_k}^t. \tag{3}$$

1. Consider first the special case $K = 1$ and $M_0 = 2$, so that $P(z) = (z - z_0)^2$. Prove that any signal of the form $s(t) = (at + b){z_0}^t$, where $a, b \in \mathbb{C}$, satisfies the recurrence equation $\forall t \in \mathbb{Z}$, $s(t) - 2z_0 s(t-1) + z_0^2 s(t-2) = 0$.

2. From now on, we are interested in the general case. Let $E$ be the set of signals $s(t)$ satisfying equation (2). Prove that $E$ is a vector space over $\mathbb{C}$.

3. Let $f$ be the map from $E$ to $\mathbb{C}^M$, which to a signal $s(t)$ associates the vector $[s(0), s(1), \ldots s(M-1)]$. Prove that the map $f$ is linear and injective (i.e. $f(s) = 0 \Rightarrow s = 0$).

4. Deduce that the dimension of $E$ is bounded by a constant and provide its expression.

5. Using equation (1), prove that $\forall k \in \{0 \ldots K-1\}$ and $\forall m \in \{0 \ldots M_k - 1\}$, $z_k$ is root of the polynomial $P^{(m)}(z) = \frac{d^m P}{dz^m}$, and check that $\forall m \geq 1$,

$$P^{(m)}(z) = \textstyle\sum_{\tau=0}^{M} p_\tau \, (M-\tau) \ldots (M-\tau-m+1) z^{M-\tau-m}.$$

6. Deduce that $\forall k \in \{0 \ldots K-1\}$, $\forall n \in \{0 \ldots M_k - 1\}$ and $\forall t \in \mathbb{Z}$, $z_k$ is also root of the function $z \mapsto \sum_{\tau=0}^{M} p_\tau \, \tau^n z^{t-\tau}$.

7. For all $k \in \{0 \ldots K-1\}$ and $m \in \{0 \ldots M_k - 1\}$, prove that the signal $s_{k,m}(t) = t^m z_k^t$ belongs to $E$ (hint: one can apply equation (2) to $s_{k,m}$ and use Newton's binomial formula: $(t-\tau)^m = \sum_{n=0}^{m} (-1)^n \binom{m}{n} t^{m-n} \tau^n$).

8. Assuming that the set $\{s_{k,m}\}_{k,m}$ is linearly independent, prove that it forms a basis of $E$.

9. Deduce that any signal $s(t)$ satisfies (2) if and only if it is expressed in the form (3).

10. Propose a method to jointly estimate the poles $z_k$ and their multiplicity $M_k$ from the samples of the signal $s(t)$.

---