Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

ABEL PEREIRA DE MACEDO BORGES JUNIOR

LOW-COMPLEXITY METHODS FOR AUTOREGRESSIVE SIGNAL MODELING

Recife

2018

Abel Pereira de Macedo Borges Junior

LOW-COMPLEXITY METHODS FOR AUTOREGRESSIVE SIGNAL MODELING

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambucocomo requisito parcial à obtenção do título de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Renato J. Cintra

Recife

2018

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

B732l Borges Junior, Abel Pereira de Macedo
Low-complexity methods for autoregressive signal modeling / Abel Pereira de Macedo Borges Junior. – 2018.
85 f.: il., fig., tab.

Orientador: Renato J. Cintra.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2018.
Inclui referências.

1. Estatística. 2. Processamento de sinais. I. Cintra, Renato J. (orientador). II. Título.

310 CDD (23. ed.) UFPE- MEI 2018-128

ABEL PEREIRA DE MACEDO BORGES JUNIOR

LOW-COMPLEXITY METHODS FOR AUTOREGRESSIVE SIGNAL MODELING

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 20 de julho de 2018.

BANCA EXAMINADORA

Prof.º Leandro Chaves Rêgo
UFC

Prof.º Hélio Magalhães de Oliveira
UFPE

Prof.º André Leite Wanderley
UFPE

À Raíza, com afeto.

AGRADECIMENTOS

A Raíza, minha esposa amada e cúmplice na vida, que esteve do meu lado nos momentos bons e maus. Além de todo o incentivo, Raíza também foi minha maior parceira de pesquisa. Um dia não é pleno sem sua companhia.

A meus pais, Abel e Bibi, por toda a paciência e cuidado que sempre deram-me gratuita e abundantemente. A Deba e Tiago, meus amigos mais queridos.

Aos amigos do DE, pela convivência agradável. Em especial, agradeço a Bruna Palm e João Eudes pela amizade despretensiosa.

A Bruna e Maelyson, por sempre estarem por perto. A Marcel Santana, João Mateus e Lucas Matheus pela amizade de tanto tempo e pelas conversas sempre produtivas sobre tecnologia e *machine learning*.

Aos colegas da In Loco, pela convivência sempre instigante. Em especial, a Rodrigo Paiva pela mentoria atenciosa.

Ao Prof Ricardo Campello, de quem tive o prazer de ouvir as aulas de matemática mais lúcidas e didáticas que já presenciei.

Ao Prof Hélio Magalhães, pela orientação dedicada do início de meus estudos de mestrado, e por ter me apresentado o problema de análise automática de ECGs.

Ao Prof Renato J. Cintra, pela orientação presente deste trabalho. O Prof Renato sempre esteve disposto a discutir assuntos relacionados à pesquisa científica e sua prática dedicada e intelectualmente honesta. Além disso, considero particularmente marcantes as conversas e *crash courses* gratuitos sobre economia.

Aos membros da banca, por investirem seu tempo na crítica deste texto.

À Valéria Bittencourt, pela solicitude de sempre.

A CAPES, pelo apoio financeiro.

Ao Criador, a quem pertence todo o conhecimento verdadeiro.

ABSTRACT

Autoregressive (AR) models provide a means to approximate the spectrum of a signal. In this work, we face the problem of designing computationally efficient methods for parameter estimation in 1st and 2nd order AR processes. First, we review how the spectral distribution provides an analysis of the variance of a time series by revealing its frequency components. Then, we tackle the low-complexity parameter estimation problem in the AR(1) case using a binarized process and a piecewise linear curve approximation heuristic, whose multiplicative complexity does not depend on the blocklength. A comprehensive literature review on the binarized version of AR(1) processes is presented. An algorithm based on stochastic approximations is presented for estimating the parameters of AR(1) processes. We show that the resulting estimator is asymptotically equivalent to the exact maximum likelihood estimator. For moderately large samples (N > 100), the algorithm represents an economy of 50% in both additions and multiplications with respect to the direct method. For the AR(2) model, based on simulations, we show how estimates of its parameters can be obtained using two iterations of AR(1) filtering. We bootstrap our AR(1) methods to solve the low-complexity AR(2) parameter estimation problem. Such iterative estimation strategy displays competitive statistical behavior in simulations when compared to standard maximum likelihood estimates. Finally, the low-complexity estimator is experimented in the context of image segmentation. The autocorrelation of pixel intensity values of texture images is considered as a descriptive measure for textures. The low-complexity estimator has a smaller within variance than the exact estimator in 30% of the considered textures and a smaller within median absolute deviation in 46% of of the cases.

Keywords: Signal Processing. Autoregressive Filters. Approximate Computing.

RESUMO

Modelos Autorregressivos (AR) provêm meios para aproximar o espectro de um sinal. Neste trabalho, abordamos o problema de desenvolver métodos computacionalmente eficientes para estimação de parâmetros de processos AR de primeira e segunda ordens. Primeiramente, revisamos como a distribuição espectral fornece uma análise da variância de uma série temporal ao revelar seus componentes de frequência. Em seguida, o problema de estimação do parâmetro de correlação do modelo AR(1) é abordado usando um processo binário e uma heurística para aproximação de curvas por uma função linear por partes. O estimador resultante tem complexidade multiplicativa independente do tamanho de bloco, N. A literatura sobre técnicas de binarização para análise de processos AR(1) é revisada. Um segundo algoritmo baseado em aproximações estocásticas é apresentado para a estimação dos parâmetros de correlação e variância de processos AR(1). Mostramos que o estimador resultante é assintoticamente equivalente ao estimador de máxima verossimilhança. Para tamanhos de amostra moderados ou grandes ($N > 100$), o algoritmo representa uma economia de 50% em adições e multiplicações relativamente ao método direto. Para processos AR(2), a partir de simulações, mostramos como seus parâmetros podem ser estimados com duas iterações de filtragem AR(1). Daí, aplicamos o estimador aproximado desenvolvido para estimação em processos AR(1) para estimar parâmetros autorregressivos de processos AR(2). Esta técnica mostra-se competitiva em simulações de Monte Carlo quando comparada com o método de máxima verossimilhança. Finalmente, o estimador aproximado é experimentado no contexto de segmentação de imagens. A autocorrelação da intensidade de *pixels* em imagens de texturas é considerada como uma medida descritiva para texturas. O estimador de baixa complexidade apresentou menor variância por grupo em 30% das 13 texturas consideradas e menor desvio absoluto mediano por grupo em 46% dos casos.

Palavras-chave: Processamento de Sinais. Filtros Autorregressivos. Computação Aproximada.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

SUMÁRIO

1 INTRODUCTION

1.1 MOTIVATION

In this work, we deal with the parameter estimation problem under computational constraints. In particular, we study a class of stationary processes called autoregressive processes of order 1, or AR(1) processes. The main motivation for this choice came from the field of image processing: the image compression problem. For motivation, we consider the very informative counting exercise presented in (GONZALEZ; WOODS, 2007, page 547). A video is a sequence of image frames. The modern high-definition television (HDTV) standard adopts frames (which are just matrices of pixel intensity values) with at least 1280-by-720 pixels. Color videos use 3 matrices in order to compose each frame, which sums up to $3 \times 1280 \times 720$ pixels. Pixel intensities are represented as 8-bit integers. Therefore, using 1 byte $= 8$ bits, the representation of a single frame needs $3 \times 1280 \times 720 > 2.7 \times 10^6$ bytes ≈ 2.7 MB of storage space. If a video is displayed at a rate of 30 frames per second, which is half the usual rate used by modern video-games, then 30 frames/second $\times$ 2.7 MB/frame $=$ 81 MB/second of data is consumed. For instance, a two-hour movie occupies approximately

$$2 \text{ hours} \times 3600 \text{ seconds/hour} \times 81 \text{ MB/second} = 583,200 \text{ MB} \approx 583.2 \text{ GB}$$

of storage space. How can we save hundreds of movies using only $1,000$ GB of storage space? The answer to this question is in image compression.

One of the most widely used tools for image compression is the Discrete Cosine Transform (DCT) (AHMED; NATARAJAN; RAO, 1974; RAO; YIP, 2014). It is present in image and video compression formats such as JPEG (WALLACE, 1992), MPEG (GALL, 1992), H.261 (International Telecommunication Union, 1990), H.263 (International Telecommunication Union, 1995), H.264/AVC (LUTHRA; SULLIVAN; WIEGAND, 2003) and HEVC (POURAZAD *et al.*, 2012). It is possible to show that the DCT matrix is the asymptotic case of the Principal Component Analysis (PCA) (JOLLIFFE, 2002) of AR(1) processes when the parameter ρ of the AR(1) process tends to 1 (AHMED; NATARAJAN; RAO, 1974, Figure 2). Therefore, the DCT works in a way analogous to the PCA: the input data is projected onto the linear span of the DCT columns and only the most significant coefficients are retained. That is the essence of (lossy) image compression. Because of this relationship, the highly-correlated AR(1) process is a commonly accepted model for the local pattern of pixel intensity values of natural images (PRATT, 2007). The main benefit of using the DCT is that the transformation

matrix is fixed: it does not depend on the input data and it works as a compressor of a fairly general class of signals.

We took a step back to study computationally efficient ways to estimate the correlation parameter ρ of autoregressive processes of order 1. Such fast estimation methods can be used to adaptively choose transformation matrices which better match the local statistics of different image regions (RADÜNZ; BAYER; CINTRA, 2016) or to approximate the correlator of synthesis-imaging arrays (ROMNEY, 1999).

In fact, it is possible to link the autoregressive parameters estimated iteratively with the frequencies in the spectrum of the signal under analysis (LI; KEDEM, 1994). That opens the door for solving common spectral analysis tasks such as signal detection and frequency estimation via autoregressive modeling techniques. We give a step towards a low-complexity approach to that class of problems in Section 5.1.2 by approximating the parameters of $\mathrm{AR}(2)$ processes.

In systhesis-imaging arrays, the correlator subsystem is responsible for the actual measurement of the interference patterns which describe the phenomena being sensed, analogously to the lens of a digital camera (ROMNEY, 1999). In particular, the FX correlator makes use of the convolution property of Fourier transforms (OPPENHEIM, 1999; SMITH, 2007) in order to compute the cross-correlation between the signals under analysis. It has two parts: the F part consists of computing the Fourier transform of the input signals, and the X part consists of computing their elementwise product. Its performance depends heavily on the architecture of complex multiply-and-accumulate (CMAC) processors, due to the use of Fourier transforms. For instance, in (LAPSHEV; HASAN, 2017), a strategy for optimizing the computation of the FX correlator is developed based on multiple CMACs. The motivation was to reduce the number of memory readings in the X part. The authors report a 30% reduction in memory consumption and 1.4% reduction in energy consumption.

The goal of the correlator is in fact to compute the cross-correlation between two signals. That can also be realized in the time-domain (ROMNEY, 1999, Section 5). Autoregressive models provide approximations for describing arbitrary signals (DJURIC *et al.*, 1999). Finally, the methods emphasized in this dissertation can provide correlator architectures with virtually zero multiplications.

1.2 OBJECTIVES

We have the following objectives:

- Review theory and methods of stochastic processes described as linear filters, in particular autoregressive filters;
- Review methods for parameter estimation in AR(1) processes;
- Propose methods for parameter estimation in AR(1) processes under hard computational constraints.

1.3 DOCUMENT STRUCTURE

The dissertation has the following structure:

- In **Chapter 2**, we briefly review concepts and literature on low-complexity algorithms used in signal processing tasks.
- In **Chapter 3**, we review the theoretical background on spectral analysis of stationary processes. In particular, we focus on how the information contained in the power spectral density furnishes valuable insights on the behavior of a stochastic process.
- In **Chapter 4**, we first use the tools of Chapter 3 to describe the statistical characteristics of a given AR(1) process y_n. We study classical estimators and we propose an approximate algorithm to estimate the variance of the process which is asymptotically equivalent to the maximum likelihood estimator and costs 50% less in terms of arithmetical complexity. Then, we consider a binary process b_n which contains only the sign information of y_n. We discuss the link between the stochastic structure of b_n and y_n under a more general assumption than the one which is usually made, which we refer to as the *symmetric assumption* (SA). We conjecture that there exists a retrieval mechanism for ρ based solely on the information from b_n when the SA is satisfied. We propose a low-complexity estimator for ρ based on a piecewise linear approximation to the map between estimators of b_n and y_n known to exist in the gaussian case. The proposed approximating function has only dyadic rational coefficients and thus its computation has null multiplicative complexity.
- Finally, in **Chapter 5**, we consider the application of the developments of Chapter 4 in two situations. The first one is the parameter estimation problem in AR(2) processes. Monte Carlo simulations suggest that we can estimate the parameters of an AR(2) process with two iterations of AR(1) filtering. This way, we "bootstrap" the estimators studied in

Chapter 4 and experiment with them in this method. The second application is in image segmentation. We carried out computational experiments with texture images and the goal was to distinguish the textures using only the first autocorrelation.

2 LOW-COMPLEXITY METHODS IN SIGNAL PROCESSING

The mathematical theory of a method for signal analysis is usually developed before an efficient way to the physical realization of the method is known to exist. This is a natural path: the concept precedes the optimization of the computation. In other words, we know *what* we want to know before we know *how* to acquire the desired knowledge. The appeal of a signal processing method depends upon factors such as what physical phenomena the signals represent and the amount of information which can be extracted from those signals by using the method.

In this chapter, we provide a brief introduction to the computational issues faced by researchers and practitioners of signal processing. We cite a few signal processing methods, or algorithms, which have proven to be valuable by providing practical insights into a wide variety of real-world signals.

2.1 COMPUTATIONAL COMPLEXITY

An algorithm is a detailed, complete description for the realization of a well-defined task. In signal processing, one of the the tasks is a computational method for processing data, i.e., for extracting information from raw signals. More specifically, the implementation of signal processing methods in digital computers needs to deal with the fact that data comes in discrete pieces. More concretely, let us say that we have an N-point input data vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\top$ and a quantity $Q(\mathbf{x})$ is to be computed. The quantity $Q(\mathbf{x})$ may be itself another data vector, or a scalar. There may exist more than one way to compute $Q(\mathbf{x})$. The computational complexity of an algorithm is the amount of computational effort spent during its execution. Amongst the factors which determine the performance of an algorithm are the hardware (or software) architecture chosen for implementation, memory consumption and the number of arithmetic operations it realizes (BRIGGS *et al.*, 1995, Section 10.6). Such factors are used to decide which algorithm–within a certain class of algorithms for computing $Q(\mathbf{x})$–is better suited for a particular application.

Given the great number of different criteria which can be used to optimize the computation of $Q(\mathbf{x})$, usually the following approximation is used:

$$\text{Computational Complexity} \approx \text{Arithmetical Complexity}. \tag{2.1}$$

The arithmetical complexity of an algorithm is measured by the number of arithmetic operations needed to complete the algorithm. Arithmetic operations include addition, multiplication, and

bit-shifting. Considering the binary representation of a number, multiplications by powers of 2 are equivalent to bit-shifting operations. Thus, we can write (2.1) in more detail as

$$\text{Computational Complexity} \approx \text{Multiplications} + \text{Additions} + \text{Bit-shifts}. \tag{2.2}$$

As the size of the problem scales up, i.e., as N grows, the number of multiplications becomes relatively more important as a measure of computational complexity in (2.2) (CINTRA; OLIVEIRA, 2015, Table 1), (FOG, 2011, page 272). For that reason, substantial research effort has been spent on optimizing (i.e., minimizing) the number of multiplications required to compute a given quantity.

2.2 FAST ALGORITHMS

We start with a "toy example" from (BLAHUT, 2010, page 2). Let $x = a + jb$ and $y = c + jd$ be complex numbers, where $j \triangleq \sqrt{-1}$ and a, b, c, d are real scalars. The multiplication xy results in the complex number $z = xy = e + jf$, where

$$\begin{cases} e &= ac - bd, \\ f &= ad + bc. \end{cases} \tag{2.3}$$

The algorithm which computes z using these equations is called the *direct method* to compute z. In this case, the direct method uses 4 multiplications and 2 additions. However, notice that we can write (2.3) equivalently as

$$\begin{cases} e &= (a-b)d + a(c-d) \\ f &= (a-b)d + b(c+d). \end{cases} \tag{2.4}$$

This new representation induces the following algorithm for computing z: (i) compute $c_1 = (a-b)d$, $c_2 = (c-d)$ and $c_3 = c+d$, then (ii) $e = c_1 + ac_2$ and $f = c_1 + bc_3$. This algorithm requires 3 multiplications and 5 additions. It trades 1 addition by 1 multiplication. Whenever additions are cheaper than multiplications, we say that this is a *fast algorithm* for the computation of z.

A fast algorithm is a procedure for the computation of a quantity Q which is more efficient than the direct method implied by the conceptual definition of Q. In Chapter 2 of his classical book, *Arithmetical Complexity of Computations*, Winograd consider three examples of algorithms which revolutionized the way many computations are performed (WINOGRAD, 1980, Chapter 2). The first one is the Toom-Cook algorithm for the multiplication of two

integers (TOOM, 1963; COOK; AANDERAA, 1969). The second one is the algorithm known as the Fast Fourier Transform (FFT), or the Cooley-Tukey FFT algorithm, for the computation of the Discrete Fourier Transform (DFT) (COOLEY; TUKEY, 1965). The third one is an algorithm for matrix multiplication discovered by Strassen in 1969 (STRASSEN, 1969).

In particular, amongst the discrete transforms, the DFT has received special attention. In (HEIDEMAN, 1988, Chapter 5), Heideman worked out a theory for the multiplicative complexity of the DFT providing closed-form lower bounds on the number of multiplications for computing an N-point DFT, as a function of N. The connection of the DFT with other discrete transforms such as the Discrete Hartley Transform (DHT) (BRACEWELL, 1983), and the DCT (AHMED; NATARAJAN; RAO, 1974) allows Heideman's theory to be extended to such cases (HEIDEMAN, 1988, Sections 6.4, 6.5).

2.3 APPROXIMATE ALGORITHMS

The research effort of the fast algorithms community is mainly focused on the exact computation of a given quantity relevant in signal processing and data analysis. We can think of such efforts as a fight against nature itself towards the cheapest way of physically realizing an important computation. Both Gauss and Cooley & Tukey were concerned about *energy* when they devised (and re-discovered) the FFT algorithm (HEIDEMAN, 1988, page 77). Gauss was worried about spending his own mental energy, whereas Cooley & Tukey's concern was mainly spending electrical energy and computer time. What about the trade-off between the accuracy and the cost of a computation? The answer to this question is in the paradigm of approximate, or inexact, computing (HAN; ORSHANSKY, 2013; BETZEL *et al.*, 2018; BASU *et al.*, 2018).

In the literature of discrete transforms, the seminal paper of Haweel (HAWEEL, 2001) introduced the signed DCT (SDCT) as an alternative to the exact DCT for use in image compression. The application of a discrete transform on an input vector $\mathbf{x}$ consists of taking inner products of $\mathbf{x}$ with a set of discrete basis functions $\mathbf{t}_1^\top, \mathbf{t}_2^\top, \ldots, \mathbf{t}_N^\top$. In matrix notation, we can write the transformed vector as $\mathbf{X} = \mathbf{T} \cdot \mathbf{x}$, where

$$X_k = \sum_{n=1}^{N} t_{k,n} x_n$$

is the kth element of $\mathbf{X}$ and $t_{k,n}$ is the (k,n)th entry of the discrete transform $\mathbf{T}$; i.e., $\mathbf{t}_k^\top$ is the kth

row of $\mathbf{T}$. The idea of the SDCT is to approximate $\mathbf{X}$ as $\text{sign}(\mathbf{T}) \cdot \mathbf{x}$, where

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0 \end{cases}$$

is the sign function and, when applied to a matrix, it acts entry-wise. The implication of this type of truncation is that the entries of $\text{sign}(\mathbf{T})$ belong to the set $\{0, \pm 1\}$. Therefore, it follows immediately that the computation of the matrix-vector product $\text{sign}(\mathbf{T}) \cdot \mathbf{x}$ has null multiplicative complexity. In (CINTRA; BAYER, 2011; BAYER; CINTRA, 2010), a similar strategy is applied to obtain DCT approximations based on the rounding function. Since the entries of the DCT matrix are also absolutely bounded by 1.5, the approaches also produced matrices with computationally simple entries with null multiplicative complexity. In (CINTRA, 2011), an unified approach to approximate the computation of discrete transforms with sinusoidal kernels is proposed based on dyadic rational rounding functions (BRITANAK; YIP; RAO, 2006, Section 5.4.4.3). Recently in (CINTRA *et al.*, 2018), the underlying idea of matrix approximation was formalized in a mixed integer programming model and applied to the convolutional kernels of convolutional neural networks. The main goal was to reduce the computational cost and speed-up the inference phase.

3 SPECTRAL ANALYSIS OF STATIONARY PROCESSES

In this chapter, we present a discussion on stochastic processes, linear filters driven by stationary processes, and the role of the spectral distribution on characterizing such processes. We aim at conveying to the reader the main ideas necessary to the development of our proposed methods. The focus is on theoretical aspects.

3.1 STOCHASTIC PROCESSES

A discrete-time stochastic process is a sequence $\{y_n, n \in \mathbb{Z}\}$ of random variables, which we refer to as simply y_n, defined in a common probability space $(U, \mathscr{E}, P)$, where U is called the sample space, or the universe, $\mathscr{E}$ is a σ-algebra of subsets of U, and P is a probability measure (POLLARD, 2002). An event E is any element of $\mathscr{E}$: $E \in \mathscr{E}$. The quantity $P(E)$ is called *the probability of* E; it measures "the size of E" within U. A real-valued random variable in $(U, \mathscr{E}, P)$ is a function y which maps U onto $\mathbb{R}$, the set of real numbers. The collection of pairs $(R, P(\{u \in U : y(u) \in R\}))$ for all $R \subset \mathbb{R}$ is called *the distribution of* y. The event $\{u \in U : y(u) \in R\}$ can be more succintly denoted as $y \in R$ and in order to differentiate that notation from the more formal one, we write $\Pr(y \in R)$ for the probability of y lying in R.

Given a point $u \in U$, a realization of y_n is induced, namely $\{y_n(u), n \in \mathbb{Z}\}$, also called a sample path. One can think of u as being the state of the world (FERGUSON, 2014), or the configuration of nature, which determines the specific realization $y_n(u)$. While a member of the sequence y_n is a random variable, which is a function, $y_n(u)$ is just a sequence of numbers.

For instance, consider the problem of determining whether the limit of a sequence exists as $n \to \infty$. Let us answer this question for (i) a "deterministic sequence" $x_n = \sum_{i=1}^{n} (1/2)^i$, for $n > 0$, with $x_n = 0$, for $n \leq 0$, and (ii) a stochastic process y_n. We define y_n as a "stochastic version" of x_n as follows. For $n \leq 0$, let $y_n = 0$ (with probability 1). For $n > 0$, let $y_n = y_{n-1} + x_n$ with probability $1/2$ and $y_n = y_{n-1}$ with probability $1/2$. That is, while $x_n = x_{n-1} + (1/2)^n$, i.e., x_n always adds $(1/2)^n$ to the previous value x_{n-1}, we flip a fair coin to decide whether $(1/2)^n$ is added to the current value y_{n-1} in order to determine the value y_n. An immediate consequence is that $y_n \leq x_n$ with probability 1. It is well known that $\lim_{n\to\infty} x_n = 1$. Therefore, even though we do not know for sure what is the value of $\lim_{n\to\infty} y_n$, the statement "$\lim_{n\to\infty} y_n$ is finite with probability 1" is valid, in the sense that the nature of y_n is in accordance with the use of this kind of terms; besides that, the statment indeed holds true. The types of assertions that we do about the sequences x_n and y_n are different in nature.

There are deterministic sequences with such a complex behavior that probabilistic statements may be used when one wants to have a glimpse on some property of the sequence. For instance, let $x_n = 1$ if n is a prime number and $x_n = 0$ otherwise. In order to know the value of x_n, one simply needs to check if n is a prime number. There is no uncertainty involved on that. However, as it turns out, when n grows, the computational realization of this check becomes increasingly harder. Let $\pi(n) = \sum_{i=2}^{n} x_i$ be the prime-counting function, which is also a deterministic sequence. Thus, $\pi(2) = 1, \pi(3) = \pi(4) = 2, \pi(5) = 3$ and so on. The Prime Number Theorem (WEISSTEIN, 2003) stands for a collection of results which provide approximations for $\pi(n)$ when $n \to \infty$. The theorem states that

$$\lim_{n\to\infty} \frac{\pi(n)}{n/\log(n)} = 1.$$

The harshness of the problem of checking if n is prime motivates the following exploration. Let n be an integer sampled uniformly at random from $\{n_1, n_1+1, \ldots, n_2\}$, where n_1 and n_2 are large-enough integers with $n_2 > n_1$: what is the probability of n being prime? The exact answer for this question is

$$\frac{\pi(n_2) - \pi(n_1 - 1)}{n_2 - n_1 + 1}.$$

The Prime Number Theorem gives an approximate answer to this question with a simple substitution of $\pi(n)$ by $n/\log(n)$. For instance, if we take $n_1 = 7001$ and $n_2 = 8000$, we have

$$\frac{\pi(8000) - \pi(7000)}{8000 - 7001 - 1} = \frac{107}{1000} = 0.107$$

and

$$\frac{8000/\log(8000) - 7000/\log(7000)}{8000 - 7001 - 1} \approx 0.100,$$

which represents a relative error of $\approx -6.54\%$. This is an example of a complex deterministic sequence whose analysis can benefit from probabilistic assertions.

The comment by Kolmogorov in (KOLMOGOROV, 1983) (seen in (RÊGO, 2007)) is useful here:

> *In everyday language we call random these phenomena where we cannot find a regularity allowing us to predict precisely their results. [...] Therefore, we should have distinguished between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of the probability theory).*

Hereafter, we use "stochastic processes", "time series" interchangeably. Also, we refer to the observed time series as a "data series".

3.2 STATIONARY STOCHASTIC PROCESSES

3.2.1 Strict stationarity

In order to completely characterize the behavior of a stochastic process y_n, we must know the joint distribution of any finite-dimensional slice $(y_{n_1}, \ldots, y_{n_k})$ from it. The collection of all such distributions contains all the information about the dynamics of y_n. When the distributions of $(y_{n_1}, \ldots, y_{n_k})$ and $(y_{n_1+\ell}, \ldots, y_{n_k+\ell})$ are the same for any $(n_1, \ldots, n_k)$, k and ℓ, we say that y_n is strictly stationary (BROCKWELL; DAVIS, 2002, page 15, Remark 1). In particular, for $k = 1$, strict stationarity implies that all marginal distributions of y_n are the same. This notion has mainly theoretical value. It means that the dynamics of the process does not depend upon the specific time window considered.

3.2.2 Wide-sense stationarity, or stationarity

A weaker concept of stability has found to be more relevant in practice. It is based on averages about y_n. Define $\mu_y(n) \triangleq \mathrm{E}(y_n)$ and $\Gamma_y(n,m) \triangleq \mathrm{E}\{(y_n - \mu_y(n))(y_m - \mu_y(m))\}$ as the mean and autocovariance functions of y_n (BROCKWELL; DAVIS, 2002, Definition 1.4.1). We say that y_n is weakly stationary, wide-sense stationary, covariance-stationary, or simply stationary if its mean is constant and its autocovariance function depends only on the discrete-time lag between observations (BROCKWELL; DAVIS, 2002, Definition 1.4.2). More precisely, $\mu_y(n)$ is constant and $\Gamma_y(n,m) \triangleq \gamma_y(|m-n|)$, for all m,n. Note that $\gamma_y(k)$ is an even function of k. We also assume, without loss of generality, that the processes have mean zero, i.e. $\mu_y(n) = 0$ for all n. In practice, this means that the sample mean is subtracted from the data series previously to subsequent analysis. The autocorrelation function of y_n is the normalized autocovariance

$$\rho_y(k) \triangleq \frac{\gamma_y(k)}{\gamma_y(0)}.$$

From the Cauchy-Schwartz inequality, $|\gamma_y(k)| \leq \gamma_y(0)$ for all k, which implies that $\rho_y(k)$ is always in $[-1,1]$ (BROCKWELL; DAVIS, 2013, page 26). We refer to $\sigma_y^2 \triangleq \gamma_y(0)$ as the variance of y_n. Unless otherwise specified, we consider the space of L^2-integrable, non-degenerate random variables, which means that the variance is a strictly positive real number: $0 < \sigma_y^2 < \infty$.

3.2.3 White noise (WN)

The simplest class of stationary processes are white noise (WN) processes. We say that y_n is a WN process if $\mathrm{E}(y_n) = 0$ and $\gamma_y(k) = \sigma_y^2 \delta_k$, where $\delta_k = 1$ if $k = 0$ and 0 otherwise. In particular, a sequence of independent and identically distributed (IID) random variables is a WN process; the converse is not true in general.

3.3 THE ENERGY SPECTRUM OF A DETERMINISTIC SEQUENCE

Consider a particular realization $\{y_n(u), n \in \mathbb{Z}\}$, or sample path, of a stochastic process y_n. In this section, for simplicity of notation, we denote $y_n(u)$ simply by y_n.

The z-transform of y_n is defined as (OPPENHEIM, 1999, Chapter 3)

$$Y(z) \triangleq \sum_{n=-\infty}^{\infty} y_n z^{-n}, \tag{3.1}$$

whenever the infinite sum converges. The values of z for which $Y(z)$ exists consists of an area in the complex plane referred to as the region of convergence (ROC) of $Y(z)$ (OPPENHEIM, 1999, page 96). The z-transform can be seen as the discrete-time analogue of the Laplace transform of a continuous-time function (BOYCE; DIPRIMA; HAINES, 2001, page 293). It is a linear operator which maps a sequence y_n into a function $Y(z)$.

The evaluation of $Y(z)$ over the unit circle $\{z \in \mathbb{C} : z = e^{j\omega}, \omega \in [-\pi, \pi)\}$, where $j \triangleq \sqrt{-1}$, is of special interest. It yields the discrete-time Fourier transform (DTFT) of y_n:

$$Y(e^{j\omega}) = \sum_{n=-\infty}^{\infty} y_n e^{-j\omega n}. \tag{3.2}$$

The DTFT is not a purely abstract mathematical concept. It is naturally linked to the physics and structure of linear time-invariant systems and its eigenfunctions (OPPENHEIM, 1999, Section 2.6.1). From the triangular inequality and the fact that the complex exponential has unit norm,

$$|Y(e^{j\omega})| \leq \sum_{n=-\infty}^{\infty} |y_n| \cdot |e^{-j\omega n}| = \sum_{n=-\infty}^{\infty} |y_n|.$$

Therefore, $\sum_{n=-\infty}^{\infty} |y_n| < \infty$ is a sufficient condition for the existence of the DTFT of y_n. In this case, y_n is said to be absolutely summable.

In general, the quantity z can be written in polar form as $z = r \cdot e^{j\omega}$, $r = |z| > 0$. Thus, (3.2) can be written as

$$Y(r \cdot e^{j\omega}) = \sum_{n=-\infty}^{\infty} (y_n r^{-n}) e^{-j\omega n}. \tag{3.3}$$

Equation (3.3) is the DTFT of $\{y_n r^{-n}\}$ and the sufficient condition for existence of the DTFT translates into $\sum_{n=-\infty}^{\infty} |y_n r^{-n}| < \infty$. It can be shown that the ROC of $Y(z)$ is equivalently determined by the collection of values of r for which (3.3) is well-defined, so that the ROC is either a ring or a disk: $\{z \in \mathbb{C} : r_L \leq |z| \leq r_U\}$, where $0 \leq r_L \leq r_U \leq \infty$ are lower and upper limits (OPPENHEIM, 1999, Section 3.2, Property 1). For instance, $r_L = r_U = r_0$ implies that $Y(z)$ converges only at the circle $\{z : |z| = r_0\}$; $r_U = \infty$ means that $Y(z)$ converges everywhere in the complex plane but inside a disk of radius r_L about the origin; $0 < r_L < r_U < \infty$ defines a ring-shaped ROC; and so on. Another consequence of this discussion is that the DTFT of y_n exists if, and only if, the ROC of $Y(z)$ contains the unit circle (OPPENHEIM, 1999, Section 3.2, Property 2). For example, let $y_n = \rho^n$ for $n \geq 0$ and $y_n = 0$ otherwise. In this case, $Y(z) = \sum_{n=0}^{\infty} \rho^n z^{-n}$ converges to $(1 - \rho/z)^{-1}$ whenever $|z| > |\rho|$. If $|\rho| < 1$, the ROC contains the unit circle and the DTFT is well-defined (summability of $|y_n|$ indeed requires $|\rho| < 1$) (OPPENHEIM, 1999, Example 3.1).

Given $Y(e^{j\omega})$, y_n can be recovered as (STOICA; MOSES *et al.*, 2005, Equation 1.2.3)

$$y_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(e^{j\omega}) e^{i\omega n} \, \mathrm{d}\omega, \tag{3.4}$$

the inverse DTFT. With the notation $y_n \overset{\text{DTFT}}{\longleftrightarrow} Y(e^{j\omega})$, we mean that y_n and $Y(e^{j\omega})$ form a DTFT pair. The energy spectrum of y_n is the squared magnitude of its spectrum:

$$\left\{ |Y(e^{j\omega})|^2 : \omega \in [-\pi, \pi) \right\}.$$

If y_n is also square-summable, i.e., $\sum_{n=-\infty}^{\infty} |y_n|^2 < \infty$, then (STOICA; MOSES *et al.*, 2005, Equation 1.2.5)

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\pi}^{\pi} |Y(e^{j\omega})|^2 \, \mathrm{d}\omega &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} y_n y_m e^{-j\omega(n-m)} \right\} \mathrm{d}\omega \\
&= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} y_n y_m \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-j\omega(n-m)} \, \mathrm{d}\omega \right\} \\
&= \sum_{n=-\infty}^{\infty} |y_n|^2.
\end{aligned}$$

This is an instance of the Parseval's identity, which is valid in a broader sense for vectors of a separable Hilbert space decomposed by orthogonal projection over a complete set of basis vectors (BROCKWELL; DAVIS, 2013, Theorem 2.4.2 (iv)). The Parseval's identity applied to the DTFT discloses a conceptually strong and practically useful truth. It says that $|Y(e^{j\omega})|^2$ provides the distribution of the "energy" in y_n in terms of its frequency components. Such frequency components are explicited in the synthesis formula (3.4), where we see that y_n admits

a representation as the superposition of infinitesimal complex sinusoids (OPPENHEIM, 1999, page 48).

In the process of obtaining Parseval's identity, one gets an interesting byproduct. Define

$$c_y(k) = \sum_{n=-\infty}^{\infty} y_n y_{n-k}, \quad k \in \mathbb{Z}, \tag{3.5}$$

as the "autocovariance-like" function of y_n. Under square-summability of y_n and from the Cauchy-Schwartz inequality, $c_y(k)$ is a real number, for all k. Now, notice that (STOICA; MOSES *et al.*, 2005, Equation 1.2.8)

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} c_y(k) e^{-j\omega k} &= \sum_{k=-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} y_n y_{n-k} \right) e^{-j\omega k} \\
&= \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} y_n y_{n-k} e^{-j\omega n} e^{-j\omega(k-n)} \\
&= \left(\sum_{n=-\infty}^{\infty} y_n e^{-j\omega n} \right) \cdot \left(\sum_{m=-\infty}^{\infty} y_n e^{j\omega m} \right) \\
&= Y(e^{j\omega}) \cdot Y(e^{-j\omega}) = |Y(e^{j\omega})|^2.
\end{aligned}$$

In other words, the energy spectrum of the sequence y_n can be obtained from the DTFT of its autocovariance-like function $c_y(k)$. This equation is useful in Section 3.5, where we study the relationship between the input and output autocovariance functions of linear time-invariant systems.

3.4 THE SPECTRAL DISTRIBUTION OF A STATIONARY PROCESS

Now, let $\{y_n, n \in \mathbb{Z}\}$ be a stationary stochastic process. We can not simply use the definition (3.2) and call it the spectrum of y_n, because, in general, a stationary process is not absolutely-summable with probability 1 (PORAT, 2008, page 27). Notice also that $Y(e^{j\omega})$ is a random variable in this case. However, y_n does have some well-defined averages, the first- and second-order moments, which characterize its dynamics (OPPENHEIM, 1999, page 65). An alternative notion of spectrum must take place.

The Wiener-Kintchine theorem (KOOPMANS, 1995), also known as Herglotz's theorem (PORAT, 2008, Theorem 2.9), (BROCKWELL; DAVIS, 2013, Section 4.3), implies that the autocovariance function of a stationary process admits the Fourier representation

$$\gamma_y(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\omega k} \, \mathrm{d}F_y(\omega), \tag{3.6}$$

where the right-hand side is a Riemann-Stieltjes integral and $F_y(\omega)$, $-\pi \leq \omega < \pi$, $F_y(-\pi) = 0$, is termed the *power spectral distribution* of y_n. Indeed, by Kolmogorov's existence theorem, there exists a random variable for which $F_y(\omega)/F_y(\pi)$ is the distribution function. As any generalized distribution function (BROCKWELL; DAVIS, 2002, page 115, Remark 2), F_y can be broken as $F_y = F_y^c + F_y^s$ (KEDEM, 1986, Section 2), where F_y^c is an absolutely continuous function and F_y^s is a step function, both positive, bounded by $F_y(\pi)$, right-continuous and monotonically non-decreasing[1]. If $F_y(\omega)$ is continuous for all $\omega \in [-\pi, \pi]$, i.e., $F_y^s = 0$, then there exists a continuous function f_y such that

$$F_y(\omega) = \int_{-\pi}^{\omega} f_y(t)\,\mathrm{d}t. \tag{3.7}$$

The function f_y is called the *power spectral density* (PSD) of y_n. Even when f_y does not exist as a continuous function, i.e., when there is a positive amount of power concentrated in a countable set of frequencies, the function f_y is still referred to as the PSD, making use of the Dirac delta function $\delta(\cdot)$. It is defined so that $\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x)dx = 1$. Then, for instance,

$$f_y(\omega) = \sum_{m=-\infty}^{\infty} \pi_m \delta(\omega - \omega_m)$$

is a discrete spectrum, where $\pi_m > 0$ is the power at frequency ω_m, with $\sum_{m=-\infty}^{\infty} \pi_m < \infty$. Notice that by the very definition of the Dirac delta function $\delta(\cdot)$, (3.7) remains valid in such cases, since

$$F_y(\omega) = \int_{-\pi}^{\omega} \sum_{m=-\infty}^{\infty} \pi_m \delta(t - \omega_m)\,\mathrm{d}t = \sum_{m=-\infty}^{\infty} \pi_m \int_{-\pi}^{\omega} \delta(t - \omega_m)\,\mathrm{d}t.$$

Under this formalism, we have the most commonly adopted definition for the PSD of a stationary process (PORAT, 2008, Section 2.8), (OPPENHEIM, 1999, Section 2.10), (STOICA; MOSES *et al.*, 2005, Section 1.3.1), which is simply the DTFT of its autocovariances, as show below:

$$f_y(\omega) = \sum_{k=-\infty}^{\infty} \gamma_y(k) e^{-j\omega k} \quad \overset{\text{DTFT}}{\longleftrightarrow} \quad \gamma_y(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\omega k} f_y(\omega)\,\mathrm{d}\,\omega. \tag{3.8}$$

If $\sum_{k=-\infty}^{\infty} |\gamma_y(k)| < \infty$, then f_y is a continuous function of ω (PORAT, 2008, page 28). Indeed, this is a common assumption in many results of time series analysis (e.g., see the theorems in (BROCKWELL; DAVIS, 2013, Chapter 3)). Again, even when f_y has discontinuities, which we incorporate into it using $\delta(\cdot)$, the representation (3.8) can still be used.

The PSD holds the following noteworthy properties (STOICA; MOSES *et al.*, 2005, Section 1.4):

[1] In fact, there is also a third component in such decomposition of F_y: the *sigular* component. We omit it here because it is not useful in most applications of signal processing.

1. The PSD is non-negative: it follows directly from its derivation from F_y as $f_y(\omega) = \frac{d}{d\omega}F_y(\omega)$ and the fact that F_y is a monotonically non-decreasing function;
2. The PSD is 2π-periodic (when considered over the whole real line):

$$f_y(\omega+2\pi) = \sum_{k=-\infty}^{\infty} \gamma_y(k)e^{-j(\omega+2\pi)k} = e^{-j2\pi k}f_y(\omega) = f_y(\omega).$$

   That is why we focus on the window $-\pi \leq \omega < \pi$ of length 2π;
3. The PSD is real and symmetric: since $\gamma_y(k) = \gamma_y(-k)$ and $e^{-j\omega k} + e^{j\omega k} = 2\cos(\omega k)$,

$$f_y(\omega) = \sigma_y^2 + 2\sum_{k=1}^{\infty} \gamma_y(k)\cos(\omega k) = f_y(-\omega).$$

   Because of this symmetry, f_y is often plotted only over the interval $[0,\pi]$;
4. The conceptual analogue of Parseval's identity for the energy spectrum of deterministic sequences arises by evaluating $\gamma_y(k)$ in (3.8) at $k = 0$, which yields

$$\sigma_y^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_y(\omega)\,\mathrm{d}\,\omega.$$

   This equation shows that (i) $(2\pi\sigma_y^2)^{-1}f_y(\omega)$ is a PDF with compact support $[-\pi,\pi]$ and (ii) it provides an analysis of the variance, or power, of y_n into the frequency components of its autocovariances. In this sense, $(2\pi\sigma_y^2)^{-1}F_y(\omega)$ can be interpreted as the *proportion of the variance* of y_n explained by frequencies not greater than ω (BROCKWELL; DAVIS, 2013, page 332). A similar interpretation is given to the ordered eigenvalues of a correlation matrix in the context of Principal Component Analysis (JOLLIFFE, 2002).

3.4.1 Examples

3.4.1.1 WN processes

If y_n is a WN process, then $\gamma_y(k) = \sigma_y^2\delta_k$ and its PSD is given by $f_y(\omega) = \sigma_y^2$. It has a *flat spectrum*; equivalently, an uniform spectral distribution. The process is named after the white color, which scatters evenly all wavelengths of light (PRIESTLEY, 1981, Section 1.1).

3.4.1.2 AR(1) processes

Let x_n be a WN process with variance σ_x^2. Let y_n be a stationary process driven by x_n as $y_n = \rho \cdot y_{n-1} + x_n$, where $|\rho| < 1$ is a constant. Then y_n is said to be a first-order autoregressive process (GRUNWALD; HYNDMAN; TEDESCO, 1995). This class of processes is the theme of

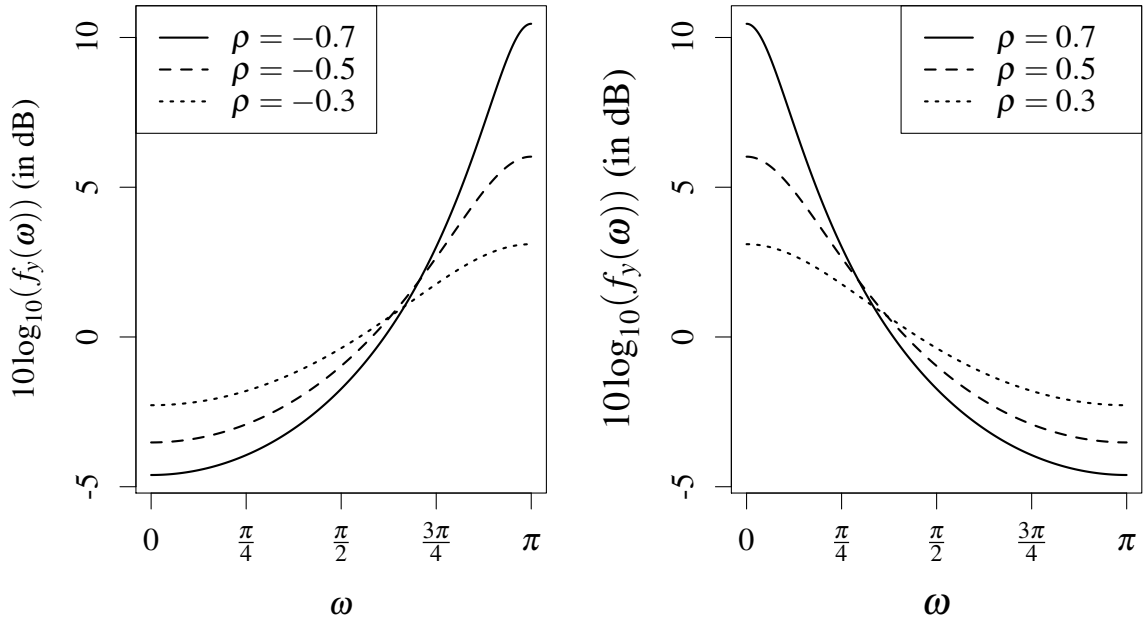

Figure 1 – Some particular cases of the PSD of the first-order autoregressive process with $\rho < 0$ (left) and $\rho > 0$ (right).

Chapter 4. We have that (see Section 4.1) (DJURIC *et al.*, 1999)

$$f_y(\omega) = \sigma_x^2 \cdot \frac{1}{\rho^2 - 2\rho\cos(\omega) + 1}.$$

Some particular cases are displayed in Figure 1, with $\sigma_x^2 = 1$. For $-1 < \rho < 0$, the process has more power given to higher frequencies. For $\rho = 0$, we have that y_n collapses into x_n and it is just a white noise with a flat spectrum. For $0 < \rho < 1$, the power is concentrated at the lower frequencies. As an illustration of this fact, two realizations of AR(1) processes are displayed in Figure 2.

3.4.1.3 MA(1) processes

Let x_n be a WN process and y_n be a stationary process obeying $y_n = x_n + \theta \cdot x_{n-1}$, where θ is a constant. This is called a first-order moving average process (BROCKWELL; DAVIS, 2002, Example x). Then $\mathrm{E}(y_n) = 0$,

$$\begin{aligned}
\gamma_y(k) &= \mathrm{E}[(x_n + \theta x_{n-1})(x_{n-k} + \theta x_{n-k-1})] \\
&= (1+\theta^2)\gamma_x(k) + \theta[\gamma_x(k+1) + \gamma_x(k-1)] \\
&= \begin{cases} (1+\theta^2)\sigma_x^2, & \text{if } k = 0, \\ \theta\sigma_x^2, & \text{if } |k| = 1, \\ 0, & \text{if } |k| > 1, \end{cases}
\end{aligned}$$

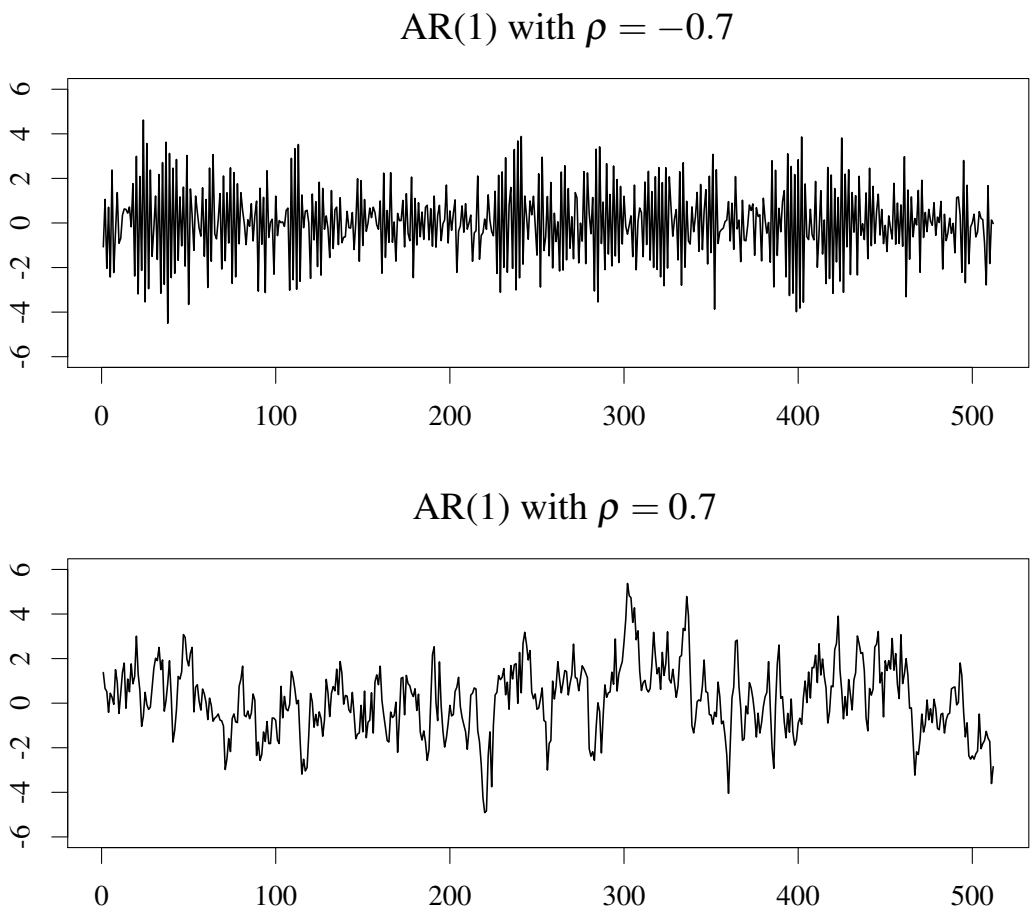

Figure 2 – Two realizations of $\mathrm{AR}(1)$ processes of length 512 with $\rho = -0.7$ (top) and $\rho = 0.7$ (bottom). Processes with $\rho < 0$ have more higher frequency components than those with $\rho > 0$. The `stats::arima.sim` function in the `R` programming environment was used with the default random number generator and seed 0.

and the PSD of y_n is given by

$$f_y(\omega) = \sigma_x^2 \cdot (\theta^2 + 2\cos(\omega)\theta + 1).$$

Some cases are displayed in Figure 3, with $\sigma_x^2 = 1$. As in the $\mathrm{AR}(1)$ processes, a phase transition occurs at $\theta = 0$, when y_n is a WN. We note, however, that the PSD is smoother than in the $\mathrm{AR}(1)$ case.

In the examples discussed so far, the stationary processes consist of purely indeterministic sequences only, using the terms of Wold's decomposition (BROCKWELL; DAVIS, 2013, Theorem 5.7.1). Because of that, $F_y^s = 0$ and the spectral density is indeed a continuous function. In the next three examples, we build a process with F_y having both a discrete part, with many jumps, and a continuous part. The main goal is to show the role played by the linearity property of the DTFT when analyzing a process which is the sum of stationary processes.

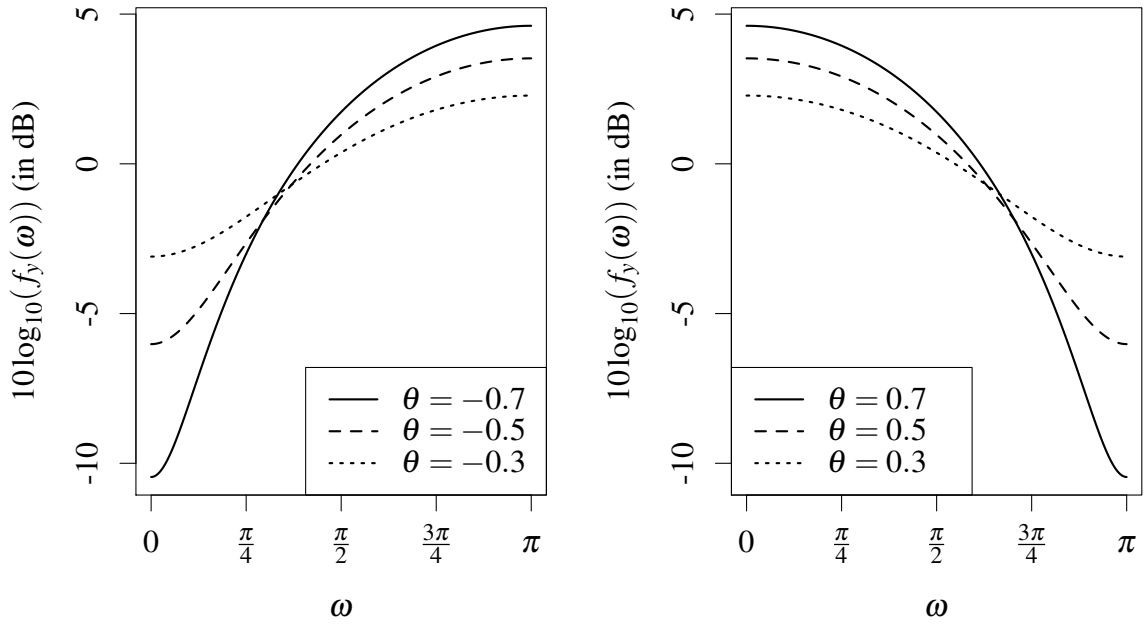

Figure 3 – Some cases of the PSD of the first-order moving average process with $\theta < 0$ (left) and $\theta > 0$ (right).

3.4.1.4 Random sinusoid

Let $y_n = A\cos(\omega_0 n) + B\sin(\omega_0 n)$, where $\omega_0 \in [0, \pi)$ is fixed and the coefficients A and B are random variables satisfying $\mathrm{E}(A) = \mathrm{E}(B) = \mathrm{E}(AB) = 0$ and $\mathrm{E}(A^2) = \mathrm{E}(B^2) = \sigma^2$. Therefore, y_n has mean zero and

$$\begin{aligned}
\mathrm{E}(y_n y_{n-k}) &= \mathrm{E}\{\ [A\cos(\omega_0 n) + B\sin(\omega_0 n)] \cdot [A\cos(\omega_0 (n-k)) + B\sin(\omega_0 (n-k))]\ \} \\
&= \sigma^2[\cos(\omega_0 n)\cos(\omega_0 (n-k)) - \sin(\omega_0 n)\sin(\omega_0 (k-n))] \\
&= \sigma^2\cos(\omega_0 k),
\end{aligned}$$

where we used the fact that the sine is an odd function and the trigonometric identities for the arc sum of cosine and sine functions. Therefore, y_n is stationary with autocovariance $\gamma_y(k) = \sigma^2\cos(\omega_0 k)$. The PSD may be written by computing the DTFT of $\gamma_y(k)$:

$$f_y(\omega) = \sigma^2\pi\{\delta(\omega + \omega_0) + \delta(\omega - \omega_0)\}, \quad -\pi \leq \omega < \pi.$$

The spectrum of y_n concentrates its power at $\pm\omega_0$. In this case, the spectral distribution $F_y = F_y^s$ is the step function defined by

$$\frac{1}{2\pi\sigma^2} \cdot F_y(\omega) \stackrel{\text{def}}{=} \frac{1}{2\pi\sigma^2} \cdot \int_{-\pi}^{\omega} f_y(t)\,\mathrm{d}t = \begin{cases} 0, & \text{if } -\pi \leq \omega < -\omega_0, \\ 1/2, & \text{if } -\omega_0 \leq \omega < \omega_0, \\ 1, & \text{if } \omega_0 \leq \omega < \pi. \end{cases}$$

3.4.1.5 Sum of random sinusoids

Let $y_n = \sum_{m=1}^{M} x_{m,n}$, where $x_{m,n} = A_m \cos(\omega_m n) + B_m \sin(\omega_m n)$, the random variables A_m and B_m are all uncorrelated with mean zero, $\mathrm{E}(A_m^2) = \mathrm{E}(B_m^2) = \sigma_m^2$, and the frequencies $0 \leq \omega_m < \pi$ are all distinct. The process $x_{m,n}$ was studied in the previous example. It is stationary with a discrete spectral distribution whose power is evenly concentrated at the frequencies $\pm\omega_m$, for a fixed value of m. Because the coefficients are all uncorrelated, it is easy to verify that $x_{m,n}$ is uncorrelated with $x_{\ell,n}$ for $\ell \neq m$. Thus, y_n is stationary and

$$\gamma_y(k) = \sum_{m=1}^{M} \gamma_{x_m}(k) = \sum_{m=1}^{M} \sigma_m^2 \cos(\omega_m k).$$

Since the DTFT is a linear transform, we have that F_y is again a step function with M jumps each of size $\frac{1}{2}\sigma_m^2$ at the $-\omega_m$s followed by other M jumps of the same size at the ω_ms.

3.4.1.6 Sum of random sinusoids plus noise

Let x_n be a stationary process with continuous PSD $f_x(\omega)$ and let y_n be the sum of random sinusoids of the previous example, uncorrelated with x_n. Then, again from the linearity of the DTFT, $z_n = y_n + x_n$ has a continuous-by-parts spectral distribution $F_z = F_z^c + F_z^s$, where the continuous part is $F_z^c = F_x$ and the discrete part is $F_z^s = F_y$. Explicitly, we have

$$F_z(\omega) = \sum_{m \in m(\omega)} \frac{\sigma_m^2}{2} + \int_{-\pi}^{\omega} f_x(t)\,\mathrm{d}t, \quad -\pi \leq \omega < \pi,$$

where $m(\omega)$ is the set of indices of the frequencies in $\{\pm\omega_m, m = 1, 2, \ldots, M\}$ which are not greater than ω and $\{\omega_m\}$ are the frequencies of the components of y_n, as in the previous example. In (KAY, 1993, Example 4.2), a variation of this example is considered within a regression framework in which the coefficients are fixed.

3.5 LINEAR TIME-INVARIANT SYSTEMS

Let $\{x_n, n \in \mathbb{Z}\}$ be a stationary time series. We will say that $\{y_n, n \in \mathbb{Z}\}$ is the output of a linear time-invariant system driven by x_n if, with probability 1,

$$y_n = \sum_{k=-\infty}^{\infty} h_k x_{n-k}, \tag{3.9}$$

for some set $\{h_k, k \in \mathbb{Z}\}$ of real numbers, called the impulse response of the filter. The name comes from the fact that, if the input is $x_n = \delta_n$, the Kronecker's delta, then the filter response is

$y_n = h_n$. The operation performed in (3.9) between h_k and x_n is a convolution. One can say that y_n is the output of the convolutional filter h_k with x_n given as input and write $y_n = (h * x)_n$.

If $h_k = 0$ for $k < 0$, y_n is a function of $\{x_m,\ m \leq n\}$ only. In this case, we say that the filter is causal. For a causal filter, if there is an integer $M > 0$ such that $h_k = 0$ for $k \geq M$, i.e., $\{h_k\}$ is a finite set and (3.9) a finite sum, we say that the filter has a finite impulse response, or it is a FIR filter; otherwise the filter has an infinite impulse response and it is an IIR filter. Motivated by operational concerns (physical computers can only perform a finite number of calculations), the problem of deriving a FIR filter which, in some sense, optimally mimics a given "ideal", target IIR filter is an important problem in signal processing (OPPENHEIM, 1999, Section 7.2).

The z-transform of h_k, $H(z)$, is called the transfer function of the filter and it plays a fundamental role in the description of its properties. If x_n and y_n are deterministic with z-transforms $X(z)$ and $Y(z)$, we have the following celebrated result.

Theorem 3.1 (Convolution Theorem (Section 2.9.6 in (OPPENHEIM, 1999))). *If $y_n = (h * x)_n$, then*

$$Y(z) = H(z) \cdot X(z)$$

for all values of z lying in the intersection of the ROCs of $X(z)$, $Y(z)$ and $H(z)$.

Note the analogue result for the Laplace transform (BOYCE; DIPRIMA; HAINES, 2001, Theorem 6.6.1) and its role in the solution of initial value problems (BOYCE; DIPRIMA; HAINES, 2001, Example 2, page 333). In particular, when these sequences all have Fourier transforms, i.e., their ROCs include the unit circle, Theorem 3.1 implies that

$$|Y(e^{j\omega})|^2 = |H(e^{j\omega})|^2 \cdot |X(e^{j\omega})|^2. \tag{3.10}$$

This equation makes clear how the operation $h * x$ change the energy spectrum of x_n. The filter h_k attenuates some frequencies and amplifies some others in the spectrum of x_n–hence the name filter.

Now, suppose that x_n is any stationary process with autocovariance function $\gamma_x(k)$, with $\sigma_x^2 = \gamma_x(0) < \infty$. Then, from (3.9) and under

$$\sum_{k=-\infty}^{\infty} |h_k| < \infty, \tag{3.11}$$

we have

$$\begin{aligned}
\mathrm{E}(y_n y_{n+k}) &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_m h_\ell \, \mathrm{E}(x_{n-\ell} x_{n+k-m}) \\
&= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_m h_\ell \gamma_x(k-(m-\ell)) \qquad\qquad (r \leftarrow m-\ell) \\
&= \sum_{r=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h_m h_{m-r} \gamma_x(k-r) \\
&= \sum_{r=-\infty}^{\infty} \gamma_x(k-r) \sum_{m=-\infty}^{\infty} h_m h_{m-r} \\
&= \sum_{r=-\infty}^{\infty} \gamma_x(k-r) c_h(r), \qquad (3.12)
\end{aligned}$$

where $c_h(r) = \sum_{m=-\infty}^{\infty} h_m h_{m-r}$ is the "autocovariance-like" function of h_n (see Equation (3.5) and the discussion around it). First of all, we see that $\gamma_y(k) = \mathrm{E}(y_n y_{n+k})$ is a function of the lag k only and, since $\mathrm{E}(y_n) = 0$, it follows that y_n is stationary. The sufficient condition here is summability of $|h_n|$. Secondly, we see that $\gamma_y(k)$ is the convolution of $\gamma_x(k)$ with $c_h(k)$. Thus, from the Convolution Theorem and the definition of the PSD, we also have a "convolution theorem" regarding the convolutional filter (3.9) with stationary inputs and outputs in terms of their PSDs and the DTFT of h_n:

$$f_y(\omega) = |H(e^{j\omega})|^2 \cdot f_x(\omega). \qquad (3.13)$$

The well-known ARMA processes (BROCKWELL; DAVIS, 2002, Chapter 3) is one of the most important classes of models for stationary processes. In this model, x_n is a WN process, and so $f_x(\omega) = \sigma_x^2$ is constant in ω. Thus, $f_y(\omega)$ essentially equals $|H(e^{j\omega})|^2$ and, in this sense, the filter coefficients completely determine the PSD of the output y_n. In other cases, when x_n is not necessarily a WN process, (3.13) is interpreted similarly to (3.10).

There are two major emphasis in the study of linear filters:

- *Filter design.* Relying on (3.10) (or (3.13)) we are faced with the problem of deriving a filter h_k which implements a given, prespecified "spectral transformation". This means that the coefficients h_k are supposed to attenuate and/or amplify some frequencies in the spectrum of x_n. Very popular applications include the processing of digital audio and musical signals (SMITH, 2007).
- *Filter estimation.* The problem is to find the filter h_k which maximizes the likelihood that $y_n = (h * x)_n$ holds true, given observations of x_n and y_n. This is sometimes called the deconvolution problem. In some cases, only y_n is observed and assumptions about the probability distribution of x_n take place in order to make the problem well-posed.

Filter estimation is usually the sense in which linear filters are introduced in statistics textbooks (BROCKWELL; DAVIS, 2002, Chapter 5). All sorts of time series analysis and forecasting can take advantage of the estimated filter.

If h_k is a function of a parameter vector $\boldsymbol{\theta}$, we have the *parametric filter*

$$y_n = \sum_{k=-\infty}^{\infty} h_k(\boldsymbol{\theta}) x_{n-k}. \tag{3.14}$$

Examples include the AR(1) and MA(1) processes discussed before. The MA(1) is already in the form of a causal filter with $h_0 = 1$, $h_1 = \theta$ and $h_k = 0$ in the other cases. Under the condition $|\rho| < 1$, the AR(1) process can also be represented as a causal filter $y_n = \sum_{k=0}^{\infty} \rho^k x_{n-k}$ (see next chapter), where we have $h_k = \rho^k$ for $k \geq 0$ and $h_k = 0$ otherwise.

Indeed, a parametric family of filters can provide models at a reduced dimension. For example, many models of practical relevance are IIR filters. In general, the characterization of these systems must be done by considering the infinite set of parameters $\{h_0, h_1, h_2, \ldots\}$. If $h_k = h_k(\boldsymbol{\theta})$ for some finite-dimensional vector $\boldsymbol{\theta}$, the analysis of the system is made simpler. Also, statistical inference problems such as estimation of the filter coefficients from a finite slice of data reduce to solving an optimization problem which is much more feasible in finite dimensions. Also, sufficient conditions on $\{h_k\}$ so that the filter is causal can be translated into constraints on $\boldsymbol{\theta}$.

4 LOW-COMPLEXITY INFERENCE FOR AR(1) PROCESSES

In this chapter, we focus on a specific subclass of the ARMA models: the AR(1) filter model (GRUNWALD; HYNDMAN; TEDESCO, 1995). From the discussion in Section 3.4.1, we can think about the predicted data series resulting from the estimated AR(1) filter as a first-order, coarse linear approximation of a signal. It is one of the simplest attempts to express the data series as an autoregression in time. In this sense, the AR(1) process is a fundamental class of time series models (DJURIC *et al.*, 1999).

Despite its simplicity, the AR(1) model is the theme of recent papers and still motivates research efforts. For instance, in (ALLÉVIUS, 2018), the inverse of the correlation matrix, the so-called precision matrix, of irregularly sampled AR(1) processes is shown to have a closed form, which is also sparse, as in the usual case of uniformly-in-time sampling (ALLÉVIUS, 2018, Equation 3). The author shows how the result allows fast pseudo-random number generation. Also, in (RESCHENHOFER, 2018), a robust estimate of the correlation coefficient, ρ, is proposed in the case of heteroscedasticity, i.e., when the variance of the series is not constant but some function of time (CRIBARI-NETO, 2004).

The concepts introduced in Chapter 3 are used here to characterize AR(1) processes. We discuss how its main parameter, ρ, provides an interpretation for the process in both time and frequency domains. We also review classical estimators for $\boldsymbol{\theta} \triangleq (\rho, \sigma_x^2)$, where σ_x^2 is the variance of the input process in the AR(1) scheme. A new approximate estimator for $\boldsymbol{\theta}$ is proposed which provides an economy of 50% in arithmetical complexity for moderate or large sample sizes ($N > 100$). Then, we focus on the development of an alternative, low-complexity estimator for ρ. Towards this goal, we study a binary threshold process derived from AR(1) data (KEDEM, 1980). Our results show that a reliable recovery of ρ under one-bit compressive sampling of AR(1) processes and computational constraints is feasible.

4.1 AR(1) PROCESSES

Let x_n be a WN process with variance σ_x^2. Let y_n be a stationary process driven by x_n as

$$y_n = \rho \cdot y_{n-1} + x_n, \quad n \in \mathbb{Z}, \tag{4.1}$$

for some real constant ρ. Then y_n is said to be an AR(1) process. This is a generative probabilistic model in the sense that the model explicitly prescribes the algorithm by which y_n is

sampled (GELMAN *et al.*, 2014, page 336). A recursive application of the defining Equations (4.1) leads to the representation of y_n as a causal filter passing through to x_n:

$$y_n = \sum_{k=0}^{\infty} \rho^k x_{n-k}. \tag{4.2}$$

We see that this equation is a particular case of the parametric linear filter (3.14) with

$$h_k = h_k(\rho) = \begin{cases} \rho^k, & \text{if } k \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Therefore, the sufficient condition (3.11) for stationarity in this case translates into $|\rho| < 1$. Thus, a stationary and causal solution y_n of (4.1) exists uniquely if $|\rho| < 1$ (PORAT, 2008).

From (3.12), since $\gamma_x(k) = \sigma_x^2 \delta_k$ and the autocovariance-like function of h_k is $c_h(r) = \sum_{\ell=-\infty}^{\infty} \rho^\ell \rho^{r+\ell} = \rho^r/(1-\rho^2)$, the autocovariance function of y_n is $\gamma_y(k) = (\gamma_x * c_h)_k = \frac{\sigma_x^2}{1-\rho^2} \cdot \rho^{|k|}$. Therefore, the variance of y_n is $\sigma_y^2 = \gamma_y(0) = \sigma_x^2/(1-\rho^2)$ and its autocorrelation function is given by the power law

$$\rho_y(k) = \rho^{|k|}, \quad k \in \mathbb{Z}. \tag{4.4}$$

From that, we have the time-domain interpretation of ρ: if $\rho \approx 1$, then consecutive observations are highly correlated and there is a high probability that they have similar values. In contrast, if $\rho = 0$, then $y_n = x_n$; thus y_n is a WN process. We note that the entire autocorrelation structure of the AR(1) process is fully characterized by the parameter ρ.

Now, let us analyze the AR(1) filter coefficients h_k defined in (4.3), with $|\rho| < 1$. The DTFT of h_k is given by the geometric series

$$H(e^{j\omega}) = \sum_{k=-\infty}^{\infty} h_k \cdot e^{-j\omega k} = \sum_{k=0}^{\infty} \rho^k \cdot e^{-j\omega k}.$$

Since $|\rho \cdot e^{-j\omega}| = |\rho| \cdot |e^{-j\omega}| = |\rho| < 1$ by hypothesis, $H(e^{j\omega})$ converges to

$$H(e^{j\omega}) = \frac{1}{1 - \rho \cdot e^{-j\omega}},$$

uniformly in ω (RUDIN *et al.*, 1976, Theorem 7.9 (Uniform Convergence Criteria)). From (4.2), it follows that the PSD of y_n is given by

$$f_y(\omega) = \sigma_x^2 \cdot \left|H(e^{j\omega})\right|^2 = \sigma_x^2 \cdot \left|\frac{1}{1 - \rho \cdot e^{-j\omega}}\right|^2 = \sigma_x^2 \cdot \frac{1}{\rho^2 - 2\rho\cos(\omega) + 1}. \tag{4.5}$$

Similarly to what happens in the autocorrelation function (4.4), the parameter ρ is the most important parameter in modeling the shape of the PSD of the AR(1) process. In fact, the

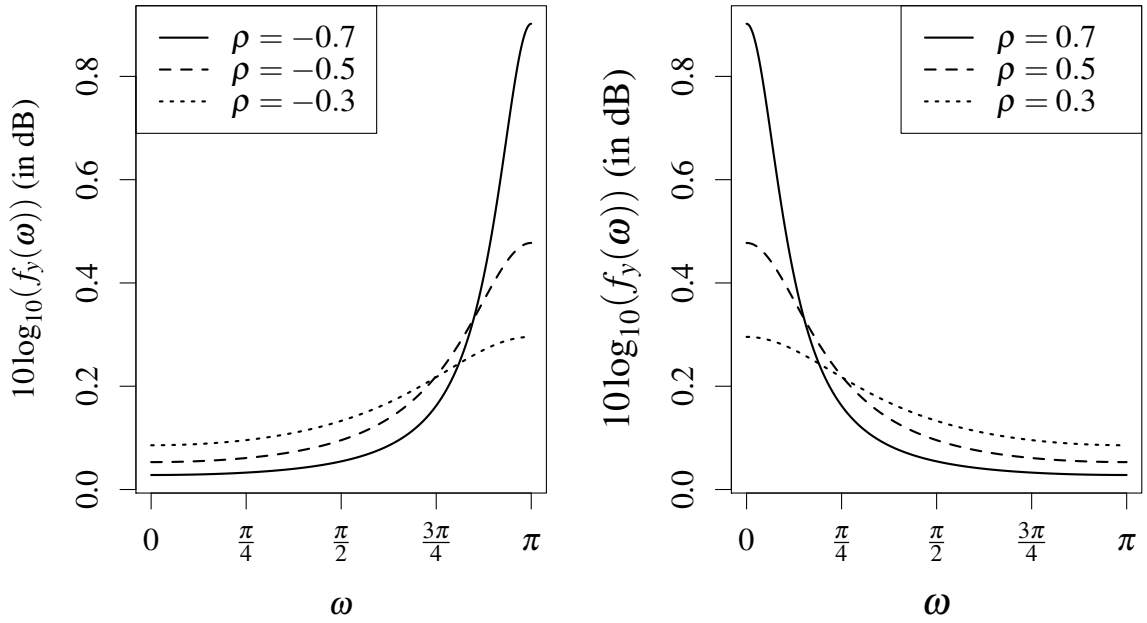

Figure 4 – Normalized PSD (4.6)**,** $0 \leq \omega \leq \pi$**, of some** AR(1) **processes. Left:** $\rho < 0$**; right:** $\rho > 0$**.**

normalized PSD, which is a PDF for $\omega \in [-\pi, \pi]$, depends only on ρ:

$$\bar{f}_y(\omega) = (2\pi\sigma_y^2)^{-1} f_y(\omega) = \frac{1}{2\pi} \cdot \frac{1-\rho^2}{\rho^2 - 2\rho\cos(\omega) + 1}, \tag{4.6}$$

For reference, we display some curves of $\bar{f}_y(\omega)$, $0 \leq \omega \leq \pi$, in Figure 4. Positive values of ρ are related to processes with low-frequency oscillations, whereas negative values of ρ indicate high-frequency behavior. This interpretation is intuitive also from a time-domain point of view. Therefore, fitting an AR(1) process to an observed data series provides a crude, glimpse-like, first-try classification of signals into low-frequency and high-frequency signals.

4.1.1 On Distributional Specifications

Up to this point, we have not examined in further details the probabilistic nature of x_n and y_n. We only assume that x_n is a WN process and y_n is stationary. In the AR(1) context, two problems can be posed (ANDĚL, 1983):

1. Given the distribution of the IID sequence x_n, find the marginal distribution of y_n.
2. Given a marginal distribution for y_n, find the distribution of the IID sequence x_n.

Specification of one of these distributions completely determine the other: they are linked by the system's defining Equations (4.1). For now, we focus on the first question above, namely how the distribution of x_n determines the common marginal distribution of y_n.

As we mentioned earlier, the equation (4.1) provides a generative model for y_n based

on the input process x_n. From that equation, the characteristic function (CF) can be used to find the marginal distribution of y_n implied by a given distribution for x_n. The CF of a random variable y is given by (MAGALHÃES, 2006)

$$\psi(t) \triangleq \mathrm{E}(\exp(jty)). \tag{4.7}$$

If y has a continuous distribution with PDF $p(y')$, then

$$\psi(t) = \int_{-\infty}^{\infty} e^{jty'} p(y')\,\mathrm{d}y'$$

is the inverse Fourier transform of $p(y')$. In the sequel, let $Q(x) \triangleq \Pr(x_n \leq x)$ be the cumulative distribution function (CDF) of x_n, $p(y)$ be the common marginal PDF of y_n, and $\psi_x(t)$ and $\psi_y(t)$ be the CF of x_n and y_n, respectively.

Since x_n is an IID sequence of random variables with common CF $\psi_x(t)$, we have that

$$\mathrm{E}(\exp\{jt(ax_n + bx_m)\}) = \mathrm{E}(\exp\{jtax_n\}) \cdot \mathrm{E}(\exp\{jtbx_m\}) = \psi_x(at) \cdot \psi_x(bt), \tag{4.8}$$

for any scalars a and b; the first identity is a direct consequence of independence and the second one follows from the definition of the CF in (4.7). From (4.2), and applying (4.8) inductively, the CF of y_n follows as (ANDĚL, 1983, Equation 2.1)

$$\psi_y(t) = \mathrm{E}\left(\exp\left\{jt\sum_{k=0}^{\infty}\rho^k x_{n-k}\right\}\right) = \prod_{k=0}^{\infty}\mathrm{E}\left(\exp\{jt\rho^k x_{n-k}\}\right) = \prod_{k=0}^{\infty}\psi_x(\rho^k t). \tag{4.9}$$

An immediate consequence is that if x_n has a symmetric distribution, then that is also the case for y_n, provided that the above infinite product converges. To see this, recall that, by definition, the CF forms a Fourier transform pair with the PDF. Since $\mathrm{E}(x_n) = 0$, the PDF of x_n must be symmetric *around zero*. Therefore, it is an even function and its (inverse) Fourier transform, which is the CF of x_n, is real. In this case, from (4.9), the CF of y_n is also real, which implies, by the same Fourier transform argument, that y_n has a symmetric marginal. We highlight this result in a lemma because it is useful in the sequel.

Lemma 4.1 (Symmetric Distributions)**.** *If the distribution of x_n is symmetric, then so is the marginal distribution of y_n.*

From now on, we use the phrase "*the symmetric assumption*" (SA) to mean that x_n and y_n are symmetrically distributed around the origin. Therefore, we have that $Q(-x) = 1 - Q(x)$ and $p(y) = p(-y)$.

4.1.2 Gaussian AR(1) Processes

If x_n is an IID sequence of gaussian random variables with mean zero and variance σ_x^2, then $\psi_x(t) = \exp\left(-\frac{t^2}{2}\cdot\sigma_x^2\right)$, which satisfies the SA and is indeed a real function. From (4.9),

$$\begin{aligned}\psi_y(t) &= \prod_{k=0}^{\infty}\exp\left(-\sigma_x^2 t^2\rho^{2k}/2\right) = \exp\left(-\frac{t^2}{2}\sigma_x^2\sum_{k=0}^{\infty}\rho^{2k}\right)\\ &= \exp\left(-\frac{t^2}{2}\cdot\frac{\sigma_x^2}{1-\rho^2}\right) = \exp\left(-\frac{t^2}{2}\cdot\sigma_y^2\right),\end{aligned}$$

which implies that y_n has gaussian marginals with mean zero and variance σ_y^2. Since correlated gaussian random variables have a joint gaussian distribution, it follows that any slice $(y_{n_1},\ldots,y_{n_k})$ from $\{y_n\}$ follows a multivariate gaussian law. In this case, y_n is said to be a Gaussian Process (KEELEY; PILLOW, 2018). In summary, this argument shows the following.

Theorem 4.1 (Gaussian Processes)**.** *If x_n consists of IID gaussians, the resulting y_n in* (4.2) *is a gaussian* AR(1) *process.*

It is noteworthy that, because a member of the gaussian family is completely determined by first- and second-order moments, stationarity leads to strict stationarity in this case. That is because both $(y_{n_1},\ldots,y_{n_k})$ and $(y_{n_1+\ell},\ldots,y_{n_k+\ell})$ are gaussian random vectors with mean zero and $\operatorname{cov}(y_{n_t},y_{n_s}) = \operatorname{cov}(y_{n_t+m},y_{n_s+m}) = \gamma_y(|n_t-n_s|)$, which means that they have the same autocovariance matrix.

4.2 BINARIZED AR(1) PROCESSES

Consider the binary threshold process $\{b_n, n\in\mathbb{Z}\}$ based on y_n given by (KEDEM; FOKIANOS, 2002, Equation 2.2)

$$b_n = \begin{cases} 1, & \text{if } y_n \geq 0,\\ 0, & \text{if } y_n < 0.\end{cases} \tag{4.10}$$

This transformation has been called hard-limiting (KEDEM, 1980; KEDEM, 1976), extreme clipping (VLECK; MIDDLETON, 1966), hard thresholding (KITIC *et al.*, 2013), amongst other nomenclatures. The analysis of b_n tends to be focused on the information content of the zero-crossings in y_n. A zero-crossing in y_n occurs from time $n-1$ to time n if y_{n-1} and y_n have different signs. A given realization of b_n is a binary sequence from which information about zero-crossings of y_n is available: if $b_{n-1}\neq b_n$, then y_{n-1} and y_n are indeed different in sign. As it

turns out, such information can be effectively used to make inferences about the original process y_n. The natural question is posed fundamentally as a concern about the information content of zero-crossings, i.e., it is an Information Theory type of question (KULLBACK, 1997; MACKAY; KAY, 2003):

> What can be inferred about an AR(1) process from its associated binary threshold process?

4.2.1 State of the Art

One can think of binarization, or one-bit sampling, as the extreme case of rounding (OPPENHEIM, 1999, Section 6.6): only one bit about each number is taken, representing the sign. Once digital computers can represent a number with only a finite-length sequence of bits, any computation with real numbers is actually an approximation (BLAHUT, 2010). The required precision varies with the context at hand.

The topic of extreme clipping, or one-bit sampling, is naturally appealing to the field of compressive sensing, where it has been indeed studied (BOUFOUNOS; BARANIUK, 2008; PLAN; VERSHYNIN, 2013). More recently, in (KIPNIS; DUCHI, 2017), the estimation of the mean of a sample of IID gaussian random variables under one-bit sampling was considered under various scenarios for the communication channel.

4.2.1.1 Motivation

The idea of investigating the information contained in level-crossings of a signal has been around for a long time as well as still being theme of recent research (SINN; KELLER, 2011; MOSSBERG; SINN, 2017). It appeals to both theory and practice of signal processing. In particular, there is an ever-growing need of algorithms for the feasible computation of fundamental quantities in signal processing responsible for a huge number of instruction calls in modern digital signal processors (DSPs) (BETZEL *et al.*, 2018). If we are able to recover useful information about a process by simple countings of its level-crossings, dramatic reductions of data storage and data processing may take place (KEDEM, 1986). We emphasize that the benefits encompass both memory usage (each b_n requires only one bit of memory) and arithmetic complexity (in the hardware level, computations involving binary sequences can be implemented as fast, lightweight procedures).

4.2.1.2 Previous work

The celebrated Rice's formula, due to the work of Stephen Rice on the mathematics of noise-corrupted signals (RICE, 1944), also known as Van Vleck's formula (VLECK; MIDDLETON, 1966, Equation 17), (JORDAN, 1986, Equation 21), (KEDEM, 1980, Equation 1.3), establishes a simple link between the covariance of a bivariate gaussian random vector (x_1, x_2) and the covariance of its signed version $(\text{sign}(x_1), \text{sign}(x_2))$, namely (JORDAN, 1986, Equation 21):

$$\text{cov}(\text{sign}(x_1), \text{sign}(x_2)) = \frac{2}{\pi} \arcsin(\text{cov}(x_1, x_2)).$$

In Section 4.3.2, we discuss this topic in further detail. This relationship has been successfully applied to time series analysis, as we show in the sequel. It turns out that the link also holds between Pearson's and Kendall's correlations of a fairly general family of elliptical distributions (LINDSKOG; MCNEIL; SCHMOCK, 2003), with a minor technical modification to include in the result special cases of distributions which concentrate probability mass in points, called atoms of the distribution. This establishes a surprising connection of zero-crossings analysis with nonparametric statistics. Interestingly, there is evidence (LINDSKOG; MCNEIL; SCHMOCK, 2003, Figure 1) that the zero-crossings estimator has less variance than the Pearson product-moment estimator for heavy-tailed data. Recently, in (MOSSBERG, 2014), the case of a general gaussian process (not necessarily an autoregressive one) was considered with applications to sensor networks. Estimators of higher-order autocorrelations of y_n are obtained from b_n based on the results in (SINN; KELLER, 2011). Even more recently, in (MOSSBERG; SINN, 2017), the work was extended to deal with bivariate gaussian processes and the cross-correlations are obtained from the respective joint zero-crossings process.

4.2.1.3 Kedem's work

An important sequence of papers in the topic of time series analysis by zero-crossings has been published since the 1970s by Benjamin Kedem. In (KEDEM, 1976), the estimation of ρ in (4.1) by using the one-bit samples (4.10) is investigated. In (KEDEM, 1980), the case of AR(p) processes is considered also under one-bit sampling. In (KEDEM, 1986), the theory of spectral analysis through the so-called higher-order crossings (KEDEM; YAKOWITZ, 1994) is elegantly exposed in a tutorial style (see (KEDEM; SLUD, 1982) for a rigorous treatment of the subject). The theory unifies previous *ad hoc* procedures of level-crossings analysis.

Kedem showed that there is a "D-domain" (in reference to the discrete-time difference operator), the domain of zero-crossings, similar to the usual spectral (Fourier) domain, where frequency information can be represented.

4.2.1.4 This work

We take the following venue:

- Research cited so far goes under the gaussian assumption. In Section 4.3, we take a step back to give a closer look at the AR(1) case under the SA only. The relationship between the stochastic structure of y_n and b_n is depicted.
- Our primary goal is the efficient estimation of $\boldsymbol{\theta} = \left(\rho, \sigma_x^2\right)^\top$.
- In Section 4.4, some standard estimators based on y_n are studied. We discuss issues of computational complexity and a fast algorithm based on an observation about common factors in the estimators for ρ and σ_x^2 is developed.
- In Section 4.5, motivated by multiplication-free computing, we consider approximating the estimator based on b_n for the gaussian case by piecewise linear functions with "computationally cheap" coefficients. This induces approximate estimators of ρ.

4.3 CHARACTERIZATION OF BINARIZED AR(1) PROCESSES

Because the AR(1) model satisfies the Markov property, the identity

$$\Pr(y_n \in A_n | y_{n-1} \in A_{n-1}, \ldots, y_{n-p} \in A_{n-p}) = \Pr(y_n \in A_n | y_{n-1} \in A_{n-1}), \quad p > 1,$$

holds true, where each A_i is an arbitrary measurable (Borel) subset of the real line. In particular, if A_i is one of the sets in $\{[0,\infty), (-\infty, 0)\}$, for all i, then it follows that b_n is a Markov chain with state space $\{0,1\}$. From (4.1), we have that

$$\Pr(b_n = 1 | y_{n-1} = y) = \Pr(x_n \geq -\rho y) = 1 - Q(-\rho y).$$

That is, $[b_n | y_{n-1} = y]$ is a Bernoulli random variable with success probability $1 - Q(-\rho y)$. Note that, from the stationarity of y_n, the transition probabilities do not depend on n, meaning that b_n is a time-homogeneous process. The process y_n is also time-homogeneous. Therefore, the transition matrix associated to b_n is given by

$$\mathbf{P} \triangleq \begin{pmatrix} P_{00} & 1 - P_{00} \\ 1 - P_{11} & P_{11} \end{pmatrix},$$

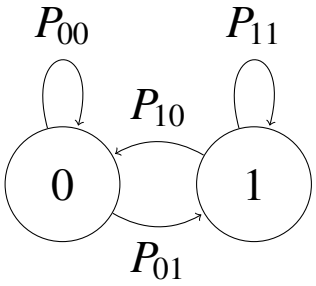

Figure 5 – Graph representation of the process b_n.

where $P_{ik} \triangleq \Pr(b_n = k | b_{n-1} = i)$. Let $\alpha \triangleq \Pr(b_n = 1)$. By the Law of Total Probability, we have

$$\begin{aligned} \Pr(b_n = 1) &= \Pr(b_n = 1, b_{n-1} = 1) + \Pr(b_n = 1, b_{n-1} = 0) \\ \therefore \quad \alpha &= \alpha P_{11} + (1-\alpha) P_{01} = \alpha P_{11} + (1-\alpha)(1 - P_{00}) \end{aligned}$$

and then P_{00} and P_{11} must satisfy

$$P_{00} = 1 - \frac{\alpha}{1-\alpha}(1 - P_{11}). \tag{4.11}$$

Therefore, the parameters α and P_{11} are sufficient to describe the process b_n. The parameter P_{00} is determined by them as in (4.11). Figure 5 displays a graph representation of the process b_n.

In the sequel, we discuss how the parameter ρ of the original process y_n determines the transition probabilities P_{ik} under the SA. This link between ρ and the elements of $\mathbf{P}$ has interesting implications on the estimation of ρ.

4.3.1 Consequences of the Symmetric Assumption (SA)

If the distribution of x_n is symmetric about zero, then $Q(-x) = 1 - Q(x)$ and

$$\Pr(b_n = 1 | y_{n-1} = y) = Q(\rho y). \tag{4.12}$$

Also, from the Lemma 4.1, y_n has symmetric marginals, which implies that $\alpha = 1/2$, and it follows from (4.11) that $P_{00} = P_{11}$. The SA gives a symmetric transition matrix to b_n, once in this case $\mathbf{P} = \mathbf{P}^\top$. Define $\lambda \triangleq P_{11} = P_{00}$. Then

$$\begin{aligned} \lambda &= \Pr(b_n = 1 | b_{n-1} = 1) \\ &= \frac{\Pr(b_n = 1, b_{n-1} = 1)}{\Pr(b_{n-1} = 1)} \\ &= 2\Pr(b_n = 1, b_{n-1} = 1) \\ &= 2\int_0^\infty \Pr(b_n = 1, y_{n-1} = y)\,\mathrm{d}y \\ &= 2\int_0^\infty \Pr(b_n = 1 | y_{n-1} = y)\Pr(y_{n-1} = y)\,\mathrm{d}y. \end{aligned} \tag{4.13}$$

The Bayes rule furnishes the second and last identities. Thus, considering (4.12), we have the following expression for λ:

$$\lambda = 2\int_0^\infty Q(\rho y)p(y)\,\mathrm{d}y. \tag{4.14}$$

That is, λ is a functional of the distributions Q and p. If Q and p belong to some parametric family of distributions, there are possibly other parameters upon which λ may depend on. Notice that, in fact, by definition,

$$P_{00} = 2\int_{-\infty}^0 (1-Q(\rho y))p(y)\,\mathrm{d}y = 2\int_0^\infty Q(\rho y)p(y)\,\mathrm{d}y = \lambda,$$

as it should hold true, according to the previous discussion. The second identity above uses once again $Q(x) = 1-Q(-x)$ and $p(y) = p(-y)$, which are consequences of the SA hypothesis. In a nutshell, so far we have learned the following:

> If y_n satisfies the AR(1) model, with $x_n \overset{\text{IID}}{\sim} Q$ and $Q(-x) = 1-Q(x)$, then the process b_n defined in (4.10) is a time-homogeneous Markov chain with state space $\{0,1\}$ and symmetric transition matrix $\mathbf{P}$ with λ and ρ linked as (4.14).

On the relationship between λ and ρ, we notice that if the correlation $\rho_y(1) = \rho$ between consecutive observations of y_n is high, then the probability

$$\lambda = P_{11} = \Pr(y_n \geq 0 | y_{n-1} \geq 0) = \Pr(y_n \leq 0 | y_{n-1} \leq 0) = P_{00}$$

of the next observation having the same sign as the present observation will be high as well. This remark suggests the existence of a bijective map between λ and ρ.

By looking at (4.14), since Q is non-decreasing by its very definition and so $Q(\rho_1 y) \geq Q(\rho_2 y)$ whenever $\rho_1 > \rho_2$, it is tempting to say that a bijective map between λ and ρ always exists under the SA. However, using the Fourier relationship between the PDF and the CF, we have

$$p(y) = \frac{1}{2\pi}\int_{-\infty}^\infty \psi_y(t)\exp\{-jty\}\,\mathrm{d}t \tag{4.15}$$

and

$$\psi_x(t) = \int_{-\infty}^\infty q(x)\exp\{jtx\}\,\mathrm{d}x, \tag{4.16}$$

where we denoted the PDF of x_n by $q(x) = \frac{\mathrm{d}}{\mathrm{d}x}Q(x)$. Substituting (4.9) into (4.15) and then using (4.16), we have

$$\begin{aligned} p(y) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\prod_{k=0}^{\infty}\psi_x(t\rho^k)\right]\exp\{-jty\}\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\prod_{k=0}^{\infty}\int_{-\infty}^{\infty}q(x)\exp\{jt\rho^k x\}\,\mathrm{d}x\right]\exp\{-jty\}\,\mathrm{d}t, \end{aligned}$$

from which is clear that $p(y)$ may depend on ρ. Thus, we can express the relationship (4.14) between λ and ρ in the more explicit form

$$\lambda = \frac{1}{\pi}\int_0^{\infty}Q(\rho y)\left\{\int_{-\infty}^{\infty}\left[\prod_{k=0}^{\infty}\int_{-\infty}^{\infty}q(x)\exp\{jt\rho^k x\}\,\mathrm{d}x\right]\exp\{-jty\}\,\mathrm{d}t\right\}\,\mathrm{d}y. \tag{4.17}$$

In order to further investigate such relationship, we consider two examples: the Cauchy and gaussian distributions. Both are symmetric distributions around the origin. However, the gaussian distribution possesses well-defined moments and is fully characterized by the first and second moments. On the other hand, the Cauchy distribution does not have any finite moment.

- *Gaussian case*: It can be shown that $\lambda = 1/2 + \pi^{-1}\arcsin\rho$ (KEDEM, 1980), which is a monotone function in the interval $\rho \in [-1,1]$. Its inverse is $\rho = \cos(\pi(1-\lambda))$. This is an important case because of the widely used gaussian assumption in time series modeling. We discuss it in details in the next section.
- *Cauchy case*: If x_n is an IID sequence of standard Cauchy random variables, then $Q(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$ and the CF is $\psi_x(t) = e^{-|t|}$. From (4.9), we have that $\psi_y(t)$ converges to $e^{-(1-|\rho|)^{-1}|t|}$ uniformly in t, which means that y_n has Cauchy marginals with scale parameter $c = (1-|\rho|)^{-1}$ and so $p(y) = \pi^{-1}c/(y^2+c^2)$. Substituting Q and p into (4.14) and calculating the integral numerically, we obtain the relationship between ρ and λ whose graphic is in Figure 6. In Figure 6, the curve of the Cauchy case is depicted along with the curve for the gaussian case and the straight line $\rho = 2\lambda - 1$ for reference.

We were not able to prove that the bijective link always exists under the SA. Based on Figure 6, we conjecture that this is indeed the case. The important implication of such conjecture being true is that it would then be possible to recover ρ after hard-limiting y_n, when we have access to the sign information b_n only, because any bijective map has an inverse.

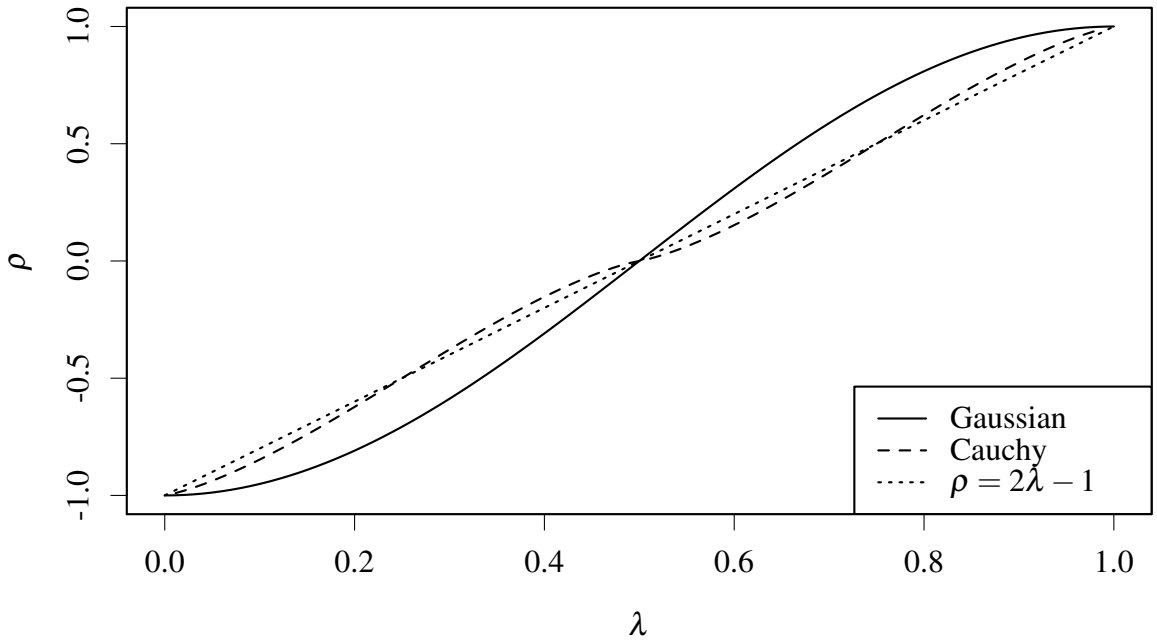

Figure 6 – Function (4.14) **computed for the gaussian case and the Cauchy case.**

4.3.2 Gaussian Inputs: Van Vleck's Formula

If y_n is a gaussian AR(1) process, the discussion in Section 4.1.2 leading to the Lemma 4.1 implies that $Q(x) = \Phi(x/\sigma_x)$ and $p(y) = \sigma_y^{-1}\phi(y/\sigma_y)$, where

$$\phi(y) = (2\pi)^{-1/2}\exp(-y^2/2) \qquad \text{and} \qquad \Phi(x) = \int_{-\infty}^{x}\phi(u)\,\mathrm{d}u = \frac{1}{2}\left\{1+\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right\}$$

are the PDF and CDF of the gaussian distribution, respectively, and $\operatorname{erf}(\cdot)$ is the error function (NG; GELLER, 1969). In this particular case of (4.14),

$$\begin{aligned}
\lambda &= \frac{2}{\sigma_y}\int_0^\infty \Phi(\rho y/\sigma_x)\phi(y/\sigma_y)\,\mathrm{d}x \qquad\qquad (y \leftarrow y/\sigma_x)\\
&= \sqrt{1-\rho^2}\left[\int_0^\infty \phi(y\sqrt{1-\rho^2})dy + \int_0^\infty \operatorname{erf}\left(\frac{\rho y}{\sqrt{2}}\right)\phi(y\sqrt{1-\rho^2})dy\right]\\
&= \frac{1}{2} + \sqrt{1-\rho^2}\int_0^\infty \operatorname{erf}\left(\frac{\rho y}{\sqrt{2}}\right)\phi(y\sqrt{1-\rho^2})dy\\
&= \begin{cases}
-\frac{1}{\pi}\tan^{-1}\left(\frac{\sqrt{1-\rho^2}}{\rho}\right), & \text{if} \quad -1<\rho<0,\\
1/2, & \text{if} \quad \rho = 0,\\
1-\frac{1}{\pi}\tan^{-1}\left(\frac{\sqrt{1-\rho^2}}{\rho}\right), & \text{if} \quad 0<\rho<1,
\end{cases}
\end{aligned} \tag{4.18}$$

where (NG; GELLER, 1969, Equation 4.3.2) furnishes the last passage. Of particular interest, the inverse link is

$$\rho = \begin{cases} -\left[1+\tan^2(\pi\lambda)\right]^{-1/2}, & \text{if } 0 \leq \lambda < 1/2, \\ 0, & \text{if } \lambda = 1/2, \\ \left[1+\tan^2(\pi\lambda)\right]^{-1/2}, & \text{if } 1/2 < \lambda \leq 1. \end{cases} \tag{4.19}$$

As mentioned before, the works reviewed in Section 4.2.1 use the fact that, under gaussianity, (y_n, y_{n-1}) follows a bivariate gaussian distribution with mean zero and covariance matrix

$$\sigma_y^2 \cdot \mathbf{R} = \frac{\sigma_x^2}{1-\rho^2} \cdot \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\mathbf{R}$ is the correlation matrix. Let $\phi(y_1, y_2; \rho, \sigma_x^2)$ be the joint PDF of (y_n, y_{n-1}). Revisiting (4.13), we have that

$$\lambda = 2\Pr(b_n = 1, b_{n-1} = 1) = 2\int_0^\infty \int_0^\infty \phi(y_1, y_2; \rho, \sigma^2)\,\mathrm{d}y_1\,\mathrm{d}y_2, \tag{4.20}$$

which is the double of the probability of (y_n, y_{n-1}) lying in the first quadrant of $\mathbb{R}^2$. Previous works are based on the evaluation of the integral in (4.20), which yields (KEDEM, 1976), (SINN; KELLER, 2011, Lemma 2.1), (GIBBONS; CHAKRABORTI, 2003, page 403)

$$\lambda = \frac{1}{2} + \frac{1}{\pi}\arcsin\rho$$

or, equivalently,

$$\rho = \sin(\pi\lambda - \pi/2) = -\cos(\pi\lambda) = \cos(\pi(1-\lambda)). \tag{4.21}$$

Notice that, in the gaussian case, the link between λ and ρ *does not depend on* σ_x^2. Equation (4.21) is known as the Van Vleck's formula (VLECK; MIDDLETON, 1966). Higher-order interactions in the form of $\mathrm{E}(b_n b_{n-k})$, for $k > 1$, have similar representations in terms of the orthant probabilities (SINN; KELLER, 2011, Section 2.1). Their evaluations require numerical integration (SINN; KELLER, 2011, Section 3.2). From the well known trigonometric identity

$$1 + \tan^2(x) = \frac{1}{\cos(x)^2},$$

it follows that the functions (4.21) and (4.19) are the same, which reaffirms our result in (4.14).

4.4 ESTIMATION OF ρ AND σ_x^2

4.4.1 Conventional Methods

Classically, the most commonly used frequentist estimators in the context of parametric probabilistic models for time series are

- the Maximum Likelihood Estimator (MLE) (BROCKWELL; DAVIS, 2013, Section 8.7),
- the Least Squares Estimator (LSE) (DJURIC *et al.*, 1999, Equation 14.92), (BROCKWELL; DAVIS, 2013, Section 8.7) and
- the Method-of-Moments Estimator (MME),

because they have well-known, desirable properties. The MME is also known as the Yule-Walker estimator (BROCKWELL; DAVIS, 2002, Section 5.1.1). In this section, we show that all these estimators are the same for ρ.

Bayesian estimators are extremely powerful, specially as a way to *represent and reason about* previously acquired information (GELMAN *et al.*, 2014, Section 1.1). Bayesian estimators also provide an MLE-like estimator, in this setting called the *maximum a posteriori* (MAP) estimator. The MAP estimator has a natural regularization feature (FRIEDMAN; HASTIE; TIBSHIRANI, 2001, Section 5.8): the prior distribution works as a regularizer (or penalizer, in the optimization jargon) for the likelihood function (FONSECA; CRIBARI-NETO, 2018, Section 3.3). In this work, we consider only frequentist estimators.

Let $\boldsymbol{\theta} = (\rho, \sigma_x^2)$ be the parameter vector to be estimated. In the frequentist framework, we assume that $\boldsymbol{\theta}$ is an unknown, constant vector. It can not be known exactly unless one has access to the whole population U of samples $u \in U$, i.e., all possible realizations of the process under study. Given a data series $\{y_1, \ldots, y_N\}$ from an AR(1) process, the MLE of $\boldsymbol{\theta}$ is given by (CASELLA; BERGER, 2002, Definition 7.2.4)

$$\bar{\boldsymbol{\theta}} \triangleq \arg\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}|y_1, \ldots, y_N), \tag{4.22}$$

where

$$\ell(\boldsymbol{\theta}|y_1, \ldots, y_N) = \log p(y_1|\boldsymbol{\theta}) + \sum_{n=2}^{N} \log p(y_n|y_{n-1}, \boldsymbol{\theta})$$

is the log of the likelihood function of $\boldsymbol{\theta}$ and $p(y_n|y_{n-1}, \boldsymbol{\theta})$ denotes the conditional distribution of y_n given y_{n-1}, which furnishes the transition probabilties of y_n. For large N, we can drop the

marginal contribution of y_1 and maximize (RESCHENHOFER, 2018, Equation 4)

$$\ell^*(\boldsymbol{\theta}|y_1,\ldots,y_N) = \sum_{n=2}^{N} \log p(y_n|y_{n-1},\boldsymbol{\theta}). \tag{4.23}$$

In the gaussian case, since x_n is gaussian and $y_n = \rho y_{n-1} + x_n$, the distribution of y_n given y_{n-1} is gaussian with mean ρy_{n-1} and variance σ_x^2. Therefore, we have

$$\ell^*(\boldsymbol{\theta}|y_1,\ldots,y_N) = -\frac{N-1}{2}\log(2\pi\sigma_x^2) - \frac{1}{2\sigma_x^2}\sum_{n=2}^{N}(y_n - \rho y_{n-1})^2. \tag{4.24}$$

The maximization problem of ℓ^* with respect to ρ is equivalent to the minimization problem of the LSE objective, the forward prediction error $\sum_{n=2}^{N}(y_n - \rho y_{n-1})^2$. Since ℓ^* is a degree-2 polynomial in ρ with negative leading coefficient, ℓ^* is convex in ρ (CHVATAL; CHVATAL *et al.*, 1983) and the solution ρ^* of $\partial\ell^*/\partial\rho = 0$, namely

$$\rho^* = \frac{\sum_{n=2}^{N} y_n y_{n-1}}{\sum_{n=2}^{N} y_{n-1}^2},$$

is unique. The quantity ρ^* is also the LSE of ρ and is asymptotically equivalent to the exact MLE of ρ, the solution of $\partial\ell/\partial\rho = 0$. The main reason for practical usage of ρ^* is that the exact MLE has not a closed form and numerical methods must be employed (RESCHENHOFER, 2018, Section 2). It is also asymptotically equivalent to the MME $\hat{\rho}$ obtained by a simple substitution of the sample counterparts of the quantities in $\rho = \rho_y(1) = \mathrm{E}(y_n y_{n-1})/\mathrm{E}(y_n^2)$. Using the MME definition in (BROCKWELL; DAVIS, 2013, Equation 7.2.1), we have

$$\hat{\rho} = \frac{\sum_{n=2}^{N} y_n y_{n-1}}{\sum_{n=1}^{N} y_n^2}. \tag{4.25}$$

The only difference between $\hat{\rho}$ and ρ^* is the extra term y_1 in the denominator's sum. Notice that $\hat{\rho} < \rho^*$ (almost surely) and $\hat{\rho} \approx \rho^*$ for large N. We focus on $\hat{\rho}$ because there is a positive probability that $\rho^* \notin [-1,1]$ (RESCHENHOFER, 2018), while the Cauchy-Schwartz inequality guarantees that

$$\left(\sum_{n=2}^{N} y_n y_{n-1}\right)^2 \leq \left(\sum_{n=2}^{N} y_n^2\right)\left(\sum_{n=2}^{N} y_{n-1}^2\right) < \left(\sum_{n=1}^{N} y_n^2\right)^2 \quad \therefore \quad |\hat{\rho}| < 1 \quad \text{(almost surely)}.$$

Similarly, the maximization of ℓ^* with respect to σ_x^2 yields

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{n=2}^{N} \hat{x}_n^2, \tag{4.26}$$

where $\hat{x}_n^2 = (y_n - \hat{\rho} y_{n-1})^2$. We have that $\hat{\boldsymbol{\theta}} = (\hat{\rho}, \hat{\sigma}_x^2)$ is asymptotically equivalent to the exact MLE of $\boldsymbol{\theta}$, $\bar{\boldsymbol{\theta}}$, and therefore it is asymptotically unbiased and attains the Cramér-Rao

lower bound (CASELLA; BERGER, 2002, Theorem 10.1.2). In other words, $\hat{\boldsymbol{\theta}}$ has the minimal possible asymptotic variance amongst all asymptotically unbiased estimators of $\boldsymbol{\theta}$. We have (BROCKWELL; DAVIS, 2013, Theorem 7.2.2)

$$\sqrt{N}\left(\hat{\rho}-\rho\right) \xrightarrow{\mathscr{D}} \mathscr{N}\left(0,\left(1-\rho^2\right)\right), \tag{4.27}$$

where the variance is obtained by replacing $\rho(k)$ by ρ^k in Bartlett's formula (BROCKWELL; DAVIS, 2013, Equation 7.2.5). Finally, the Invariance Principle (CASELLA; BERGER, 2002, Theorem 7.2.10) ensures that the MLE of $\sigma_y^2 = \sigma_x^2/(1-\rho^2)$ is asymptotically equivalent to $\hat{\sigma}_y^2 = \hat{\sigma}_x^2/(1-\hat{\rho}^2)$.

The estimator $\hat{\boldsymbol{\theta}}$ is obtained as the approximate MLE under the gaussian assumption. Even though, $\hat{\boldsymbol{\theta}}$ is commonly used in general because of the similarity with the MME, which does not require the gaussian assumption and is asymptotically gaussian as in (4.27). The convergence result (4.27) provides a way to build confidence intervals for $\hat{\rho}$.

4.4.2 Computational Cost Analysis and an Approximate Estimator for σ_x^2

In order to compute $\hat{\rho}$ directly from (4.25), we need to perform $(N-1)+(N)+(1) = 2N$ multiplications and $(N-2)+(N-1) = 2N-3$ additions. Also, if one has access to two slots of memory at any time, then $\hat{\rho}$ can be computed passing only once through the data by simultaneously accumulating the two sums. If only one slot of memory is available, two iterations through the data are required and, after computing the first sum, (e.g., the numerator) its value must be stored.

Also, since the computation of each $\hat{x}_n^2 = (y_n - \hat{\rho}\, y_{n-1})^2$ requires 1 addition and 2 multiplications, direct computation of $\hat{\sigma}_x^2$ in (4.26) requires $2(N-1)+1 = 2N-1$ multiplications (one less if $N-1$ is chosen to be a power of 2) and $(N-1)+(N-2)+(1) = 2N-2$ additions (one less if the value $N-1$ is some sort of universal constant in the computing system which does not need to be computed).

Thus, direct computation of $\hat{\boldsymbol{\theta}}$ as in (4.25) and (4.26) requires $(2N)+(2N-1) = 4N-1$ multiplications and $(2N-3)+(2N-2) = 4N-5$ additions.

Now, let $S \triangleq \sum_{n=1}^{N} y_n^2$ so that $\hat{\rho} = S^{-1} \sum_{n=2}^{N} y_n y_{n-1}$. Notice that

$$\begin{aligned}
\sum_{n=2}^{N} \hat{x}_n^2 &= \sum_{n=2}^{N} y_n^2 + \hat{\rho}^2 \sum_{n=2}^{N} y_{n-1}^2 - 2\hat{\rho} \sum_{n=2}^{N} y_n y_{n-1} \\
&= S - y_1^2 + \hat{\rho}^2 (S - y_N^2) - 2\hat{\rho} \cdot \hat{\rho} S \\
&= \left[S + \hat{\rho}^2 S - 2\hat{\rho}^2 S \right] - \left[y_1^2 + \hat{\rho}^2 y_N^2 \right] \\
&= S\left(1 - \hat{\rho}^2\right) - R, \qquad (4.28)
\end{aligned}$$

where $R \triangleq y_1^2 + \hat{\rho}^2 y_N^2$. Let $\hat{s}_x^2 \triangleq S\left(1 - \hat{\rho}^2\right)/(N-1)$. We have the following result.

Proposition 4.1. *The statistic $\hat{s}_x^2$ converges in probability to $\hat{\sigma}_x^2$.*

Demonstração. We have that $(N-1)\,\mathrm{E}\left\{\left|\hat{s}_x^2 - \hat{\sigma}_x^2\right|\right\} = \mathrm{E}\{|R|\}$ and, from the triangular inequality, $\mathrm{E}(|R|) \leq \mathrm{E}(y_1^2) + \mathrm{E}(\hat{\rho}^2 y_N^2)$. Hölder's inequality yields

$$\mathrm{E}(\hat{\rho}^2 y_N^2) \leq \sqrt{\mathrm{E}(\hat{\rho}^4)} \cdot \sqrt{\mathrm{E}(y_N^4)}. \qquad (4.29)$$

The fourth moment of the gaussian distribution is given by

$$\mathrm{E}\left\{\mathcal{N}(\mu, \sigma^2)^4\right\} = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4. \qquad (4.30)$$

From the asymptotic distribution of $\hat{\rho}$ in (4.27), we have that $\mathrm{E}(\hat{\rho}) \to \rho$ and

$$\begin{aligned}
\mathrm{var}\left\{\sqrt{N} \cdot (\hat{\rho} - \rho)\right\} &\to 1 - \rho^2 \\
\therefore \quad \mathrm{var}\,(\hat{\rho}) &\sim \frac{1-\rho^2}{N} \\
\therefore \quad \mathrm{E}(\hat{\rho}^4) &\sim \rho^4 + 6\rho^2 \frac{1-\rho^2}{N} + 3\frac{\left(1-\rho^2\right)^2}{N^2},
\end{aligned}$$

where "$x \sim y$" is meant to be interpreted here as "x scales as y as N grows" and the last approximation uses (4.30). From the gaussian assumption, y_N is distributed as $\mathcal{N}(0, \sigma_y^2)$ and then (4.30) yields

$$\mathrm{E}(y_N^4) = 3\sigma_y^4.$$

By applying the limit in N on both sides of (4.29), we get

$$\lim_{N\to\infty} \mathrm{E}(\hat{\rho}^2 y_N^2) \leq \sqrt{3}\rho^2\sigma_y^2.$$

Also from the gaussian assumption, $\mathrm{E}(y_1^4) = 3\sigma_y^4$. Thus

$$\lim_{N\to\infty} \mathrm{E}(|R|) \leq 3\sigma_y^4 + \sqrt{3}\rho^2\sigma_y^2 = \text{constant in } N,$$

Estimator	Method	$\mathscr{M}$	$\mathscr{A}$
$\hat{\rho}$	Diret	$2N$	$2N-3$
$\hat{\sigma}_x^2$	Direct	$2N-1$	$2N-2$
$\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \bar{\boldsymbol{\theta}}$	Direct	$4N-1$	$4N-5$
$\left(\hat{\rho}, \hat{s}_x^2\right) \xrightarrow{\text{p}} \bar{\boldsymbol{\theta}}$	Algorithm 1	$2N+3$	$2N-1$

Table 1 – Arithmetic complexity for estimators of $\boldsymbol{\theta}$ as a function of the sample size N.

and we have

$$\lim_{N\to\infty} \mathrm{E}\left\{\left|\hat{s}_x^2 - \hat{\sigma}_x^2\right|\right\} = \lim_{N\to\infty} \frac{1}{N-1}\mathrm{E}(|R|) = 0.$$

Now, from Markov's inequality, for any $\varepsilon > 0$, it holds true that

$$\lim_{N\to\infty} \Pr\left\{\left|\hat{s}_x^2 - \hat{\sigma}_x^2\right| > \varepsilon\right\} \leq \frac{1}{\varepsilon} \cdot \lim_{N\to\infty} \mathrm{E}\left\{\left|\hat{s}_x^2 - \hat{\sigma}_x^2\right|\right\} = 0.$$

We proved that $\hat{s}_x^2 \xrightarrow{\text{p}} \hat{\sigma}_x^2$, where $\xrightarrow{\text{p}}$ means convergence in probability. □

We propose Algorithm 1 for the approximate estimation of $\boldsymbol{\theta}$. It computes $\hat{\rho}$ and $\hat{s}_x^2$. Based on Proposition 4.1 and the discussion about $\hat{\rho}$ in the previous subsection, the Algorithm 1 yields an estimator with the same nice asymptotic properties as the exact MLE $\bar{\boldsymbol{\theta}}$. It requires $2N+3$ multiplications and $2N-1$ additions. As $N \to \infty$, define

$$\text{Ratio}_{\mathscr{M}}(N) \triangleq \frac{2N+3}{4N-1} \downarrow \frac{1}{2} \quad \text{and} \quad \text{Ratio}_{\mathscr{A}}(N) \triangleq \frac{2N-1}{4N-5} \downarrow \frac{1}{2}$$

as the ratios between the arithmetic complexities of $(\hat{\rho}, \hat{s}_x^2)^\top$ and $\hat{\boldsymbol{\theta}}$, where $a_n \downarrow a$ means that a_n is a monotonically decreasing sequence in n which converges to a. Therefore, we can say that, asymptotically, Algorithm 1 provides an economy of 50% in arithmetic complexity relatively to $\hat{\boldsymbol{\theta}}$. In Figure 7, we display these ratios as a function of N. It is also clear in Algorithm 1 that only two slots of memory are required if the data comes into the computing system via a data stream in an on-line fashion; otherwise it would need space in memory and N additional slots are required. In Table 1, we give the arithmetic complexity for the computation of $\hat{\rho}$ and $\hat{\sigma}_x^2$ separately, of $\hat{\boldsymbol{\theta}}$ as the sum of these two complexities, and of $\left(\hat{\rho}, \hat{s}_x^2\right)^\top$ through Algorithm 1.

We display signal-flow diagram representations of the first part of Algorithm 1 in Figure 8. The diagram in Figure 8 implements the for-loop and yields intermediary values of $\hat{s}_x^2$ and $\hat{\rho}$. We display it separately in order to make clear that the second part is not run after every new observations comes into the computing system as it is the case with the intermediary values of the parameters.

Algorithm 1: Algorithm for the Computation of $\left(\hat{\rho}, \hat{s}_x^2\right)^\top$

Require: $y_1, y_2, \ldots, y_N$
Ensure: $\hat{\rho}$ and $\hat{s}_x^2$

$\hat{s}_x^2 \leftarrow y_1^2$
$\hat{\rho} \leftarrow 0$
for $n \leftarrow 2, 3, \ldots, N$ **do**
  $\hat{s}_x^2 \leftarrow \hat{s}_x^2 + y_n^2$
  $\hat{\rho} \leftarrow \hat{\rho} + y_n y_{n-1}$
end for
$\hat{\rho} \leftarrow \hat{\rho} / \hat{s}_x^2$
$\hat{s}_x^2 \leftarrow \hat{s}_x^2 (1 - \hat{\rho}^2)/(N-1)$

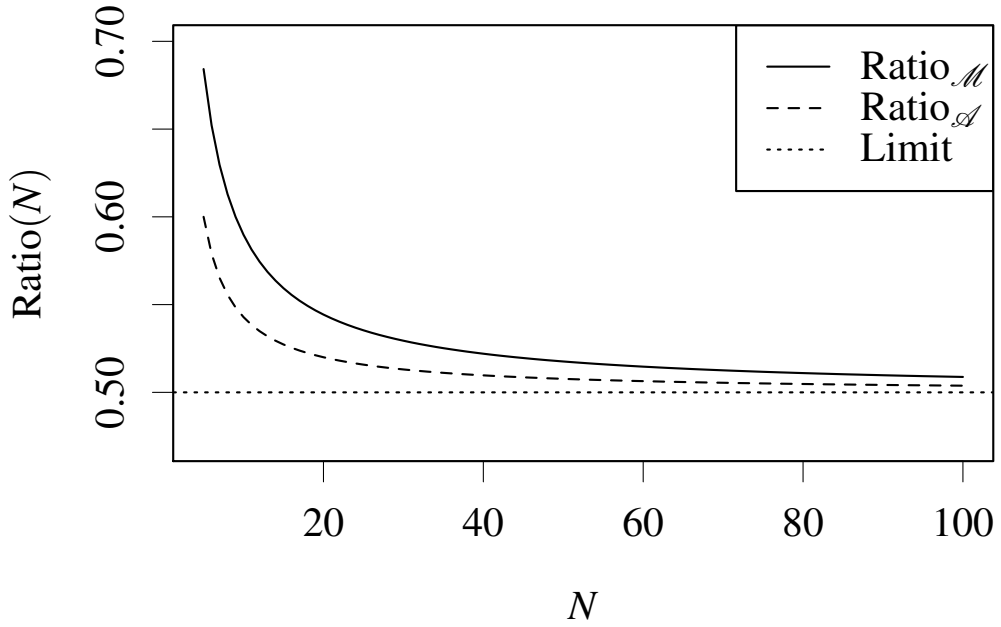

Figure 7 – Decay of the ratio between the arithmetic complexities of the direct and proposed fast implementation to estimate $\boldsymbol{\theta}$.

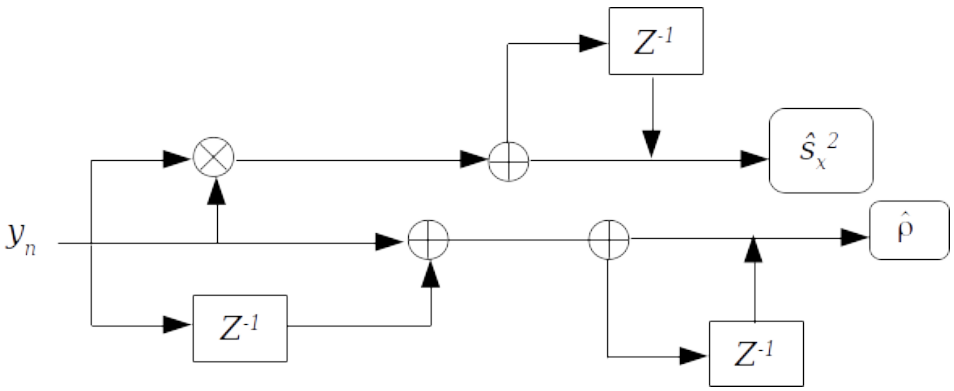

Figure 8 – First part of the signal-flow diagram representation of Algorithm 1. Here, $\hat{s}_x^2$ and $\hat{\rho}$ denote the intermediary values available after the for-loop in Algorithm 1. The notation Z^{-1} represents a time delay. The second part only concludes the final two lines in Algorithm 1.

4.4.3 Estimation of ρ Based on b_n

Let $y_1, y_2, \ldots, y_N$ be a N-point slice of contiguous elements from the AR(1) process y_n. Suppose that only one-bit of each y_n can be captured as b_n in (4.10). The problem now translates into working out statistical inference for Markov chains and using that to make inferences about the original process y_n. Statistical inference for Markov chains is a very important topic in applied statistics and it has been considered for general Markov processes in, e.g., (BILLINGSLEY, 1961) and (ANDERSON; GOODMAN, 1957). We consider only the case of interest: two-state (binary), discrete-time Markov chains as treated in Klotz's paper (KLOTZ, 1973). He actually treats the case when $\alpha = \Pr(b_n = 1)$ is not necessarily $1/2$; we use $\alpha = 1/2$.

Under the SA and assuming that y_1 follows the stationary distribution of y_n in (4.9), we have that $\alpha = \Pr(b_1 = 1) = 1/2 = \Pr(b_1 = 0)$. An intuitive estimator for λ is the empirical proportion of times when the chain keeps its state (0-0 or 1-1 transitions):

$$\hat{\lambda} = \frac{1}{N-1} \#\{n = 2, 3, \ldots, N : b_n = b_{n-1}\}, \tag{4.31}$$

where $\#A$ is the number of elements in the set A. We now check that $\hat{\lambda}$ is in fact the MLE for λ and consider its properties and the implications in the estimation of ρ.

The probability that the observed sequence equals a specific binary pattern

$$(b_1, b_2, \ldots, b_N) \in \{0,1\}^N$$

follows from the Markov property and the Bayes rule as (KLOTZ, 1973, Equation 2.6)

$$\begin{aligned}
\Pr(b_1, b_2, \ldots, b_N) &= \Pr(b_1) \prod_{n=2}^{N} \Pr(b_n | b_{n-1}) \\
&= \frac{1}{2} \prod_{n=2}^{N} \lambda^{b_n b_{n-1} + (1-b_n)(1-b_{n-1})} (1-\lambda)^{b_n(1-b_{n-1}) + (1-b_n) b_{n-1}} \\
&= \frac{1}{2} \lambda^{(N-1) - [2(t_1 - t_{11}) - t_0]} (1-\lambda)^{2(t_1 - t_{11}) - t_0}
\end{aligned} \tag{4.32}$$

where $t_1 = \sum_{n=1}^{N} b_n$ is the number of 1s, $t_{11} = \sum_{n=2}^{N} b_n b_{n-1}$ is the number of times that both b_n and b_{n-1} equal 1, and $t_0 = b_1 + b_N$. According to the Factorization Theorem (CASELLA; BERGER, 2002, Theorem 6.2.6), $D = 2(t_1 - t_{11}) - t_0$ is a sufficient statistic for λ, since we can write the likelihood $\Pr(b_1, b_2, \ldots, b_N)$ as a function of λ and D only (CASELLA; BERGER, 2002, Definition 6.2.1). This means that D efficiently summarizes the information that the sample $b_1, b_2, \ldots, b_N$ gives about λ.

We have that D is the number of zero-crossings in $y_1, \ldots, y_N$, i.e., the number of times when $b_n \neq b_{n-1}$, for $n = 2, 3, \ldots, N$ (KEDEM, 1980). The MLE of λ, say $\hat{\lambda}$, is the maximizer

of (4.32). Standard optimization routines lead to

$$\hat{\lambda} = 1 - \frac{D}{N-1}.$$

This is in accordance with our intuition in (4.31): $N-1-D$ is the number of times when the chain keeps its state. Let $k_n = 1$ if $b_n = b_{n-1}$ and $k_n = 0$. Then k_n indicates when the chain b_n keeps its state and, similarly, $1-k_n$ indicates a zero-crossing at the time $n-1$ to time n. We can write

$$\hat{\lambda} = \frac{1}{N-1}\sum_{n=2}^{N} k_n. \tag{4.33}$$

Since $\mathrm{E}(k_n) = \Pr(k_n = 1) = \Pr(b_n = b_{n-1})$, we have

$$\begin{aligned}
\mathrm{E}(k_n) &= \Pr(b_n = 0, b_{n-1} = 0) + \Pr(b_n = 1, b_{n-1} = 1) \\
&= \Pr(b_{n-1} = 0)P_{00} + \Pr(b_{n-1} = 1)P_{11} \\
&= \frac{\lambda}{2} + \frac{\lambda}{2} = \lambda.
\end{aligned}$$

Therefore, $\mathrm{E}(\hat{\lambda}) = \lambda$, i.e., $\hat{\lambda}$ is an unbiased estimator for λ. Its variance is given by

$$\begin{aligned}
(N-1)^2 \operatorname{var}(\hat{\lambda}) &= \sum_{n=2}^{N} \operatorname{var}(k_n) + 2\sum_{n=2}^{N}\sum_{m=n+1}^{N} \operatorname{cov}(k_n, k_m) \\
&= (N-1)\lambda(1-\lambda) + 2\sum_{n=2}^{N}\sum_{m=n+1}^{N} \operatorname{cov}(k_n, k_m),
\end{aligned}$$

where we used that $\operatorname{var}(k_n) = \mathrm{E}(k_n^2) - \mathrm{E}(k_n)^2 = \lambda(1-\lambda)$. For $r > 0$,

$$\begin{aligned}
\operatorname{cov}(k_n, k_{n+r}) &= \mathrm{E}(k_n k_{n+r}) - \mathrm{E}(k_n)\,\mathrm{E}(k_{n+r}) \\
&= \Pr(k_n = 1, k_{n+r} = 1) - \lambda^2 \\
&= \Pr(k_{n+r} = 1 | k_n = 1)\Pr(k_n = 1) - \lambda^2 \\
&= \lambda \Pr(k_{n+r} = 1 | k_n = 1) - \lambda^2 \\
&= \lambda \Pr(b_{n+r} = b_{n+r-1} | b_n = b_{n-1}) - \lambda^2.
\end{aligned}$$

From the time-homogeneity of b_n, $\operatorname{cov}(k_n, k_{n+r})$ does not depend on n. Let $K_r = \Pr(b_{n+r} = b_{n+r-1} | b_n = b_{n-1})$. Then (SINN; KELLER, 2011, Equation 2.1),

$$\begin{aligned}
\operatorname{var}(\hat{\lambda}) &= \frac{\lambda(1-\lambda)}{N-1} + \frac{2}{(N-1)^2}\sum_{r=1}^{N-1}(N-r)\operatorname{cov}(k_n, k_{n+r}) \\
&= \frac{\lambda(1-\lambda)}{N-1} + \frac{2\lambda}{(N-1)^2}\sum_{r=1}^{N-1}(N-r)(K_r - \lambda).
\end{aligned} \tag{4.34}$$

The quantities K_r have no closed form. The numerical evaluation of K_r is the theme of (SINN; KELLER, 2011, Section 3). However, from (KLOTZ, 1973, Equation 4.1), we have the following convergence result for $\hat{\lambda}$:

$$\sqrt{N}(\hat{\lambda}-\lambda) \xrightarrow{D} \mathcal{N}(0, 2\lambda(1-\lambda)). \tag{4.35}$$

From the Invariance Principle, we obtain an approximate MLE for ρ from (4.14) as the quantity $\hat{\rho}_a$ which solves $\hat{\lambda} = 2\int_0^\infty Q(\hat{\rho}_a y)p(y)dy$. In the gaussian case (4.21), we have $\hat{\rho}_a = \cos(\pi(1-\hat{\lambda}))$ and, for standard Cauchy inputs, $\hat{\rho}_a \approx 2\hat{\lambda}-1$ (see Figure 6). From that, we define

$$\hat{\rho}_a \triangleq \cos\left(\pi(1-\hat{\lambda})\right) \tag{4.36}$$

and we refer to $\hat{\rho}_a$ as the *approximate estimator* of ρ.

4.5 LOW-COMPLEXITY ESTIMATION OF ρ

In this section, our goal is to propose a low-complexity estimator for ρ based on the process b_n. We denote this estimator by $\tilde{\rho}$ and we call it the *low-complexity* estimator, whereas $\hat{\rho}$ and $\hat{\rho}_a$ are called the *exact* and *approximate* estimators, respectively. The estimator $\tilde{\rho}$ is obtained by approximating the function $g(\lambda) = \cos(\pi(1-\lambda))$ which links λ and ρ in the gaussian case in (4.14). We start from a piecewise linear approximation of $g(\lambda)$. Then, the coefficients obtained in this step are rounded to suitable dyadic rationals.

4.5.1 Dyadic Rational Approximation of a Real Number

The set $\mathscr{D}$ of dyadic rationals consists of the rational numbers which can be written as $n/2^m$, where n and $m \geq 0$ are integers (BRITANAK; YIP; RAO, 2006, Section 5.4.4.3). The set $\mathscr{D}$ is dense in the real line, in the sense that for any real number x and any $\varepsilon > 0$, we can find a dyadic rational $d \in \mathscr{D}$ such that $|x-d| < \varepsilon$. In this very sense, for a given m, the set $\mathscr{D}_m \triangleq \{n/2^m : n \in \mathbb{Z}\}$ of mth order dyadic rationals can be seen as an "2^{-m}-scale approximation" of the real numbers. Also, $\mathscr{D}_m \subset \mathscr{D}_{m+1}$ for all $m \geq 0$. Therefore, for an arbitrary real number x, there is a sequence $\{\tilde{x}_0, \tilde{x}_1, \ldots\}$ of $\mathscr{D}_m$-approximations of x whose associated absolute error sequence $\{|\tilde{x}_0 - x|, |\tilde{x}_1 - x|, \ldots\}$ is non-increasing, i.e., $|\tilde{x}_{i+1} - x| \leq |\tilde{x}_i - x|$. Such sequence of $\mathscr{D}_m$-approximations of x can be constructed as follows.

Consider the nearest-integer round function defined as (CINTRA, 2011)

$$\text{round}(x) \triangleq \text{sign}(x) \cdot \lfloor |x| + 1/2 \rfloor,$$

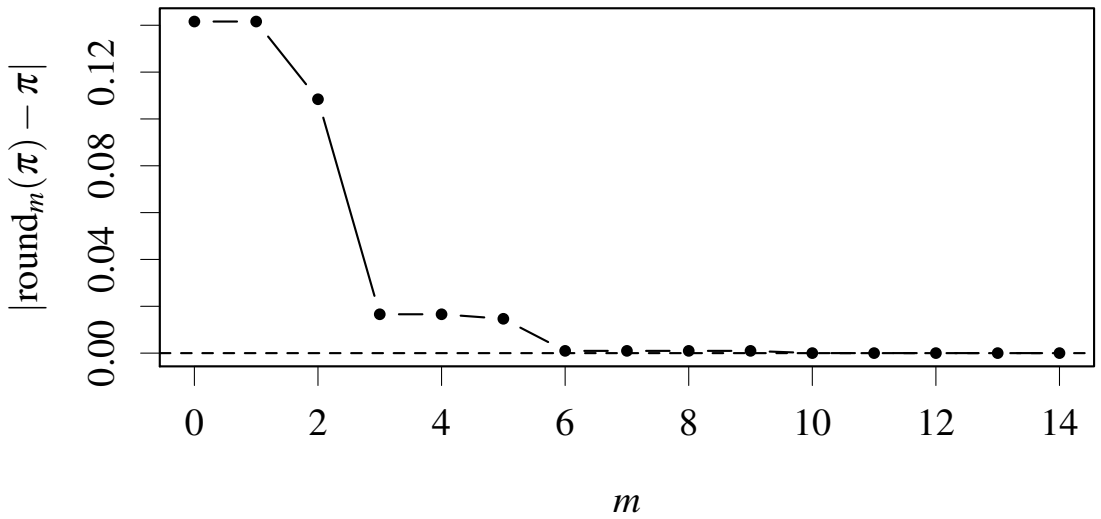

Figure 9 – The first 15 elements of the sequence of absolute errors of dyadic rounding of π.

where $\text{sign}(x)$ equals 1 if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$, and $\lfloor x \rfloor \triangleq \max\{n \in \mathbb{Z} : n \leq x\}$ is the floor of x. The image of $\text{round}(\cdot)$ is the set $\mathbb{Z}$ of integers. In fact, by construction, we have that, for any real number x,

$$\text{round}(x) = \arg\min_{n \in \mathbb{Z}} |x - n|.$$

Thus, $\text{round}(x)$ is the optimal projection (or representation) of x in the set $\mathbb{Z}$. Similarly,

$$\text{round}_m(x) \triangleq \frac{\text{round}(2^m x)}{2^m}, \quad m \geq 0, \tag{4.37}$$

maps $\mathbb{R}$ onto $\mathscr{D}_m$. In particular, $\text{round}_0(\cdot) = \text{round}(\cdot)$ and $\mathscr{D}_0 = \mathbb{Z}$. From the same reasoning, by construction, $\text{round}_m(x)$ is the optimal projection of x onto $\mathscr{D}_m$. Thus, it is clear that any real number x can be approximated arbitrarily well by some element of $\mathscr{D}_m$ by choosing a sufficiently large m. More precisely, for all x and $\varepsilon > 0$, we can always find $m(\varepsilon)$ such that $|x - \text{round}_m(x)| < \varepsilon$ for $m \geq m(\varepsilon)$. We have proven the following lemma.

Lemma 4.2. *The sequence* $\{\text{round}_m(x) : m \geq 0\}$ *converges to* x *as* $m \to \infty$.

In Figure 9, we display the absolute errors $|\text{round}_m(\pi) - \pi|$ of the dyadic rational approximation of the number π for $0 \leq m \leq 14$. For instance, $|\text{round}_6(\pi) - \pi| \approx 10^{-3}$.

4.5.2 A Piecewise Linear Curve Approximation Approach

We consider a piecewise linear curve approximation approach inspired by the method described in (HAMANN; CHEN, 1994). There are two main differences in our approach. Firstly,

we used the least squares regression to fit each straight line whereas linear interpolation is used in (HAMANN; CHEN, 1994). Secondly, because we do not use interpolation, we chose the breakpoints in a different way, based on a heuristic measure of how well the neighborhood of a given point can be approximated by straight line.

Let $I_1, I_2, \ldots, I_K$ denote a partition of $[0,1]$. That is, $I_k \bigcap I_i$ is empty whenever $k \neq i$ and $\bigcup_{k=1}^{K} I_k = [0,1]$. For each I_k we obtain the best linear approximation of $g(\lambda)$, $\lambda \in I_k$, with the ordinary least squares (OLS) linear regression method (MONTGOMERY; PECK; VINING, 2012). The input data for the regression is the set of pairs

$$\mathscr{P}_M \triangleq \left\{ \left(\frac{i}{M}, \cos\left(\pi \left[1 - \frac{i}{M} \right] \right) \right) : i = 0, 1, \ldots, M \right\},$$

for some integer $M > 1$. There are $M+1$ pairs in $\mathscr{P}_M$, which are equally spaced (in the λ direction) points lying on the graphic of the function $g(\lambda)$. Define $\mathscr{P}_M(k) \triangleq \{(\ell, r) \in \mathscr{P}_M : \ell \in I_k\}$ as the pairs of $\mathscr{P}_M$ for which the x-axis value is in I_k. Then, let $\hat{g}(\lambda)$ be the function defined by parts as $\hat{g}_k(\lambda) = c_k + d_k\lambda$, $\lambda \in I_k$, where the numbers c_k and d_k are the OLS regression estimates of the best straight line approximation of the pairs in $\mathscr{P}_M(k)$. Finally, the low-complexity approximation of $g(\lambda)$ is the function $\tilde{g}(\lambda)$ defined by parts as

$$\tilde{g}_k(\lambda) \triangleq c_k + \text{round}_{m_k}(d_k) \cdot \lambda, \quad \lambda \in I_k. \tag{4.38}$$

The order of dyadic rational rounding m_k may be different amongst the sets $I_1, I_2, \ldots, I_K$. Finally, the low-complexity estimator $\tilde{\rho}$ of ρ is given by evaluating $\tilde{g}(\lambda)$ at the MLE of λ:

$$\tilde{\rho} \triangleq \tilde{g}(\hat{\lambda}). \tag{4.39}$$

According to this construction, the function $\tilde{g}(\lambda)$ is determined by the way we cut $[0,1]$ into non-overlapping pieces, followed by the choice of the rounding orders m_k. More precisely, we first obtain $\hat{g}(\lambda)$, by choosing (i) the length K of the partition, and (ii) real numbers $t_1, t_2, \ldots, t_{K-1}$ such that $I_1 = [0, t_1)$, $I_k = [t_{k-1}, t_k)$ for $k = 2, 3, \ldots, K-1$, and $I_K = [t_{K-1}, 1]$. Note that in the case $t_k = k/K$, the intervals I_k all have the same length (in the Lebesgue sense) and the following limit holds true: $\lim_{K\to\infty} \hat{g}(\lambda) = g(\lambda)$, point-wise in λ.

Therefore, based on Lemma 4.2, the following result about the asymptotic behavior of the proposed low-complexity approximation scheme is valid.

Proposition 4.2. *If $t_k = k/K$, i.e., under uniform partitioning of $[0,1]$, the function $\tilde{g}(\lambda)$ obtained by the described method converges to $g(\lambda)$ for each $\lambda \in [0,1]$ as $K \to \infty$ and $m_k \to \infty$, $k = 1, 2, \ldots, K$.*

Our goal is to find a parsimonious low-complexity approximation $\tilde{g}(\lambda)$ of $g(\lambda)$ which is computationally cheap to calculate and whose precision suffice for most applications. Given K, we use the following algorithm in order to get insight on how to choose the breakpoints t_k:

1. For each point $p_i = \left(\frac{i}{M}, \cos\left(\pi\left[1-\frac{i}{M}\right]\right)\right)$ in $\mathscr{P}_M$ which is not an endpoint, i.e., $i \neq 0$ and $i \neq M$, find the line passing through p_{i-1} and p_i and use it to predict the y-coordinate of p_{i+1} given the x-coordinate of p_{i+1}.
2. Compute the squared error E_i between the predicted p_{i+1} y-coordinate and its actual value. We use E_i as a figure of merit to understand how well the neighborhood of p_i can be approximated with a straight line.
3. Let $E = \max_i E_i$ and define the final figure of merit of point p_i as the normalized squared error $E_i/E \in [0,1]$.

The values of E_i/E are plotted against the respective λ values in Figure 10 for the function $g(\lambda)$; we used $M = 1000$. The horizontal dashed line marks the threshold level of 0.25. The points p_i which satisfy $p_i \leq 0.25$ correspond to the interval $\lambda \in [1/3, 2/3]$. Thus, in the sense of the measure in Figure 10, the divergence from linearity of $g(\lambda) = \cos(\pi(1-\lambda))$ for $\lambda \in [1/3, 2/3]$ is at most 25% the maximal divergence, which happens at the endpoints $\lambda = 0$ and $\lambda = 1$.

From these observations, we propose the following low-complexity approximation. We used $K = 5$. The target function $g(\lambda)$ has the following anti-symmetry property:

$$g(\lambda) = \cos(\pi(1-\lambda)) = -\cos(\pi\lambda) = -g(1-\lambda). \tag{4.40}$$

The partition length $K = 5$ implies the choice of $K-1 = 4$ breakpoints. From (4.40), we constrain t_3 and t_4 as $t_3 = 1-t_2$ and $t_4 = 1-t_1$, respectively. Therefore, we only need to tweak the parameters t_1 and t_2. The same reasoning applies to any odd value of K: only $t_1, t_2, \ldots, t_{(K-1)/2}$ must be fine-tuned. We consider a "brute force" grid search approach. For each point $(t_1, t_2) \in [0.05, 0.19] \times [0.20, 0.35]$ and $m_1, m_2, m_3 \in \{0,1,2,3,4\}$:

1. Fit the piecewise linear approximation $\hat{g}(\lambda)$ of $g(\lambda)$ with breakpoints t_1, t_2, $t_3 = 1-t_2$ and $t_4 = 1-t_1$.
2. Compute $\tilde{g}(\lambda)$ from $\hat{g}(\lambda)$ as in (4.38) using m_1, m_2, m_3, $m_4 = m_2$ and $m_5 = m_1$.
3. Let the figure of merit be the supremum absolute error between $\tilde{g}(\lambda)$ and $g(\lambda)$:

$$E(t_1, t_2, m_1, m_2, m_3) = \sup_{\lambda \in [0,1]} |\tilde{g}(\lambda) - g(\lambda)|.$$

For the breakpoints' search space, a resolution of 10^{-2} was used. The minimum, with respect to $E(\cdot)$, is 0.0223. It is attained at $(t_1, t_2, m_1, m_2, m_3) = (0.14, 0.29, 4, 4, 0)$. The respective

piecewise linear curve is

$$\tilde{g}^*(\lambda) \triangleq \begin{cases} -1.0157 + \frac{11}{16}\cdot\lambda, & \text{if } \lambda \in [0.00, 0.14), \\ -1.1931 + \frac{31}{16}\cdot\lambda, & \text{if } \lambda \in [0.14, 0.29), \\ -1.5032 + 3\cdot\lambda, & \text{if } \lambda \in [0.29, 0.71), \\ -0.7602 + \frac{31}{16}\cdot\lambda, & \text{if } \lambda \in [0.71, 0.86), \\ 0.3337 + \frac{11}{16}\cdot\lambda, & \text{if } \lambda \in [0.86, 1.00]. \end{cases} \tag{4.41}$$

The quantity $\sum_{k=1}^{5} m_k = 2(m_1 + m_2) + m_3$ can be seen as a measure of the *simplicity* of $\tilde{g}(\lambda)$. If we add the constraints $m_1 = m_2 = m_3 = 0$, the best configuration is $(t_1, t_2) = (0.11, 0.33)$, which yields $E(\cdot) = 0.0587$. Such error is more than 2.6 times larger than the maximal deviation of $\tilde{g}(\lambda)$. For that reason, we use only $\tilde{g}(\lambda)$ in our experiments. We notice that $-1.0157 + \frac{11}{16}\cdot 0.01 \approx -1.0088$ and $0.3337 + \frac{11}{16}\cdot 0.99 \approx 1.0143$. In order to guarantee that the estimates of ρ are within $[-1, 1]$, we define

$$\tilde{g}(\lambda) \triangleq \begin{cases} -1, & \text{if } \tilde{g}^*(\lambda) \leq -1, \\ \tilde{g}^*(\lambda), & \text{if } \tilde{g}^*(\lambda) \in (-1, 1), \\ 1, & \text{if } \tilde{g}^*(\lambda) \geq 1. \end{cases} \tag{4.42}$$

The induced low-complexity estimator of ρ is then $\tilde{\rho} = \tilde{g}(\hat{\lambda})$. In Figure 11, we display the curves $g(\lambda)$ and $\tilde{g}(\lambda)$. In Figure 12, we display a signal-flow diagram representation of the computation of $\tilde{g}^*(\lambda)$.

4.5.3 Computational Cost Analysis

VanVleck's (or Kedem's) approximate estimator is given by

$$\hat{\rho}_a = g(\hat{\lambda}) = \cos\left(\pi(1 - \hat{\lambda})\right).$$

From $\cos(\pi(1-\lambda)) = -\cos(\pi\lambda)$, given $\hat{\lambda}$, the estimator $\hat{\rho}_a$ requires 1 multiplication $(\lambda \cdot \pi)$ and the computation of the cosine function; the result has the sign bit reversed before it is output. The CORDIC algorithm (LAKSHMI; DHAR, 2010) and its variations are the most common choice for the computation of elementary trigonometric functions at the software level when no hardware multiplier is available. Use cases include biomedical applications embedded on microcontrollers or FPGAs (KWONG; CHANDRAKASAN, 2011). It requires only additions,

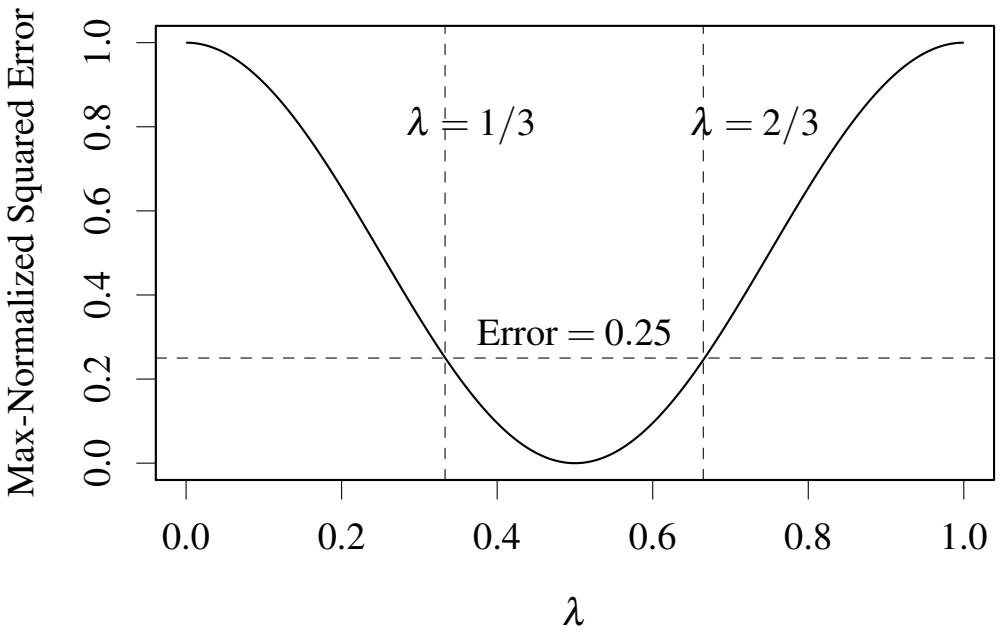

Figure 10 – Error measure E_i/E through the domain of λ for the function $g(\lambda)$.

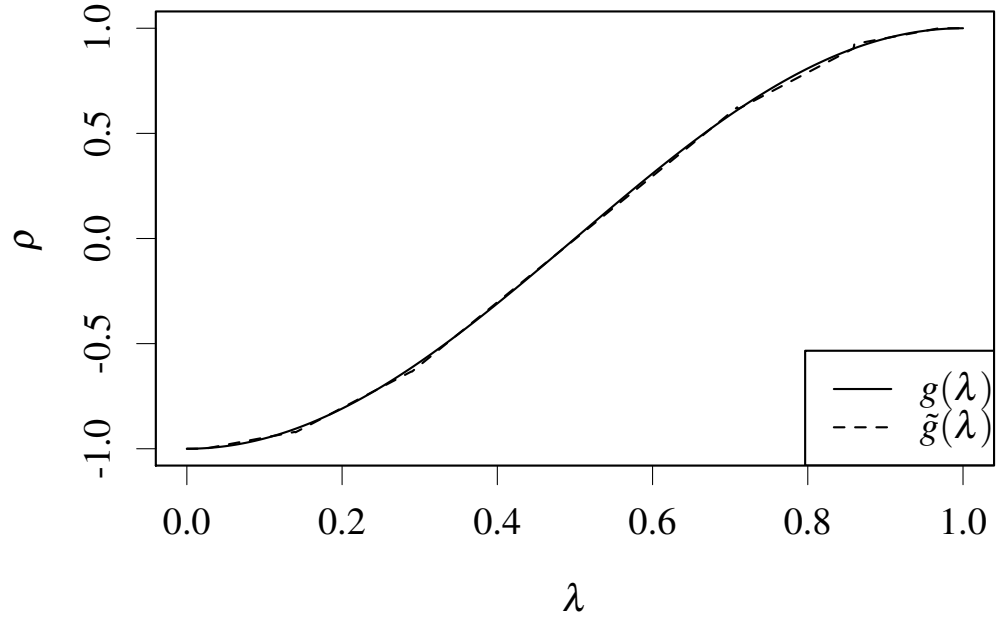

Figure 11 – Link functions from λ to ρ for the approximate ($g(\lambda)$) and low-complexity ($\tilde{g}(\lambda)$) estimators.

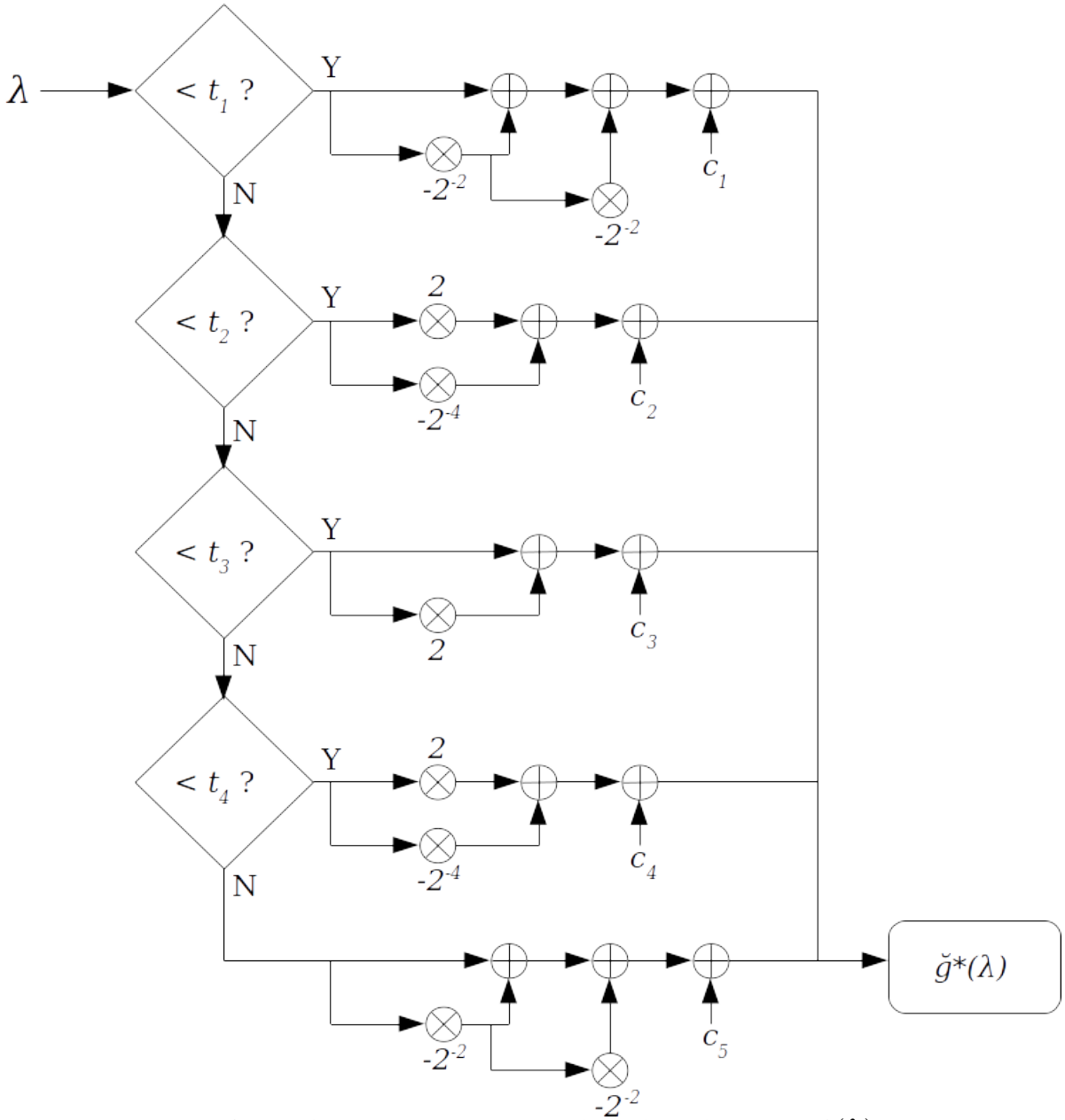

Figure 12 – Signal-flow graph representation of the function $\tilde{g}^*(\lambda)$ in (4.41).

bit-shifts and the precomputation and storage of a lookup table with floating point numbers which are used through the iterations of the algorithm.

Our alternative estimator is $\tilde{\rho} = \tilde{g}(\hat{\lambda})$, where $\tilde{g}(\cdot)$ is given in (4.42). Thus, the estimator $\tilde{\rho}$ requires the storage of $4+5=9$ floating-point numbers and 3 dyadic rationals: the breakpoints, the intercepts and the slopes. For the middle case, notice that $3\lambda = 2\lambda + \lambda$. That is, multiplication by 3 can be implemented with 1 bit-shift and 1 sum. Also, since

$$\frac{31}{16} = \frac{32-1}{16} = 2 - \frac{1}{2^4},$$

multiplication by $31/16$ can be implemented with 5 bit-shifts and 1 sum. Finally, we have that

$$\frac{11}{16} = \frac{16-(4+1)}{16} = 1 - \frac{1}{2^2} - \frac{1}{2^4}$$

and then multiplication by $11/16$ can be implemented with 6 bit-shifts and 2 sums. Notice that we could also write

$$\frac{11}{16} = \frac{8+2+1}{16} = \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^4},$$

however 8 bit-shifts are required under this representation. Therefore, in a worst-case scenario, given an estimate of λ, the estimator $\tilde{\rho}$ requires no more than 3 sums and 6 bit-shifts.

If a prior distribution $\pi(\ell)$ is available for λ, the expected value of the number of sums can be computed as

$$3\left(\int_0^{0.14} \pi(\ell)\,\mathrm{d}\ell + \int_{0.86}^{1.00} \pi(\ell)\,\mathrm{d}\ell\right) + 2\int_{0.14}^{0.86} \pi(\ell)\,\mathrm{d}\ell.$$

Similarly, the expected value of the number of bit-shifts is

$$6\left(\int_0^{0.14} \pi(\ell)\,\mathrm{d}\ell + \int_{0.86}^{1.00} \pi(\ell)\,\mathrm{d}\ell\right) + \int_{0.29}^{0.71} \pi(\ell)\,\mathrm{d}\ell + 5\left(\int_{0.14}^{0.29} \pi(\ell)\,\mathrm{d}\ell + \int_{0.71}^{0.86} \pi(\ell)\,\mathrm{d}\ell\right).$$

In particular, if $\pi(\ell) = \ell^{-1}$, $\ell \in [0,1]$, is the uniform distribution, the expected number of sums is 2.28 and the expected number of bit-shifts is 3.60.

4.6 STATISTICAL PERFORMANCE ANALYSIS

It looks like we loose a great deal of information by taking only one bit b_n of each observation y_n. For instance, since b_n says nothing about the amplitude of y_n, we can not estimate the variance σ_y^2 of the process y_n based solely on b_n. A natural question arises: what does one loose in terms of estimation accuracy when one uses $\hat{\rho}_a = \cos(\pi(1-\hat{\lambda}))$, or $\tilde{\rho} = \tilde{g}(\hat{\lambda})$ instead of $\hat{\rho}$ to estimate ρ?

In order to answer this question, we consider the mean squared error (MSE) as a figure of merit of an estimator. The MSE of $\hat{\rho}$ for a sample of size N is defined as (CASELLA; BERGER, 2002, Section 7.3)

$$\mathrm{MSE}(\hat{\rho},N) \triangleq \mathrm{E}\left\{(\hat{\rho}-\rho)^2\right\}. \tag{4.43}$$

Analogously, we have that $\mathrm{MSE}(\dot{\rho},N) \triangleq \mathrm{E}\left\{(\dot{\rho}-\rho)^2\right\}$, where $\dot{\rho} \in \{\hat{\rho}_a, \tilde{\rho}\}$ is an alternative estimator of ρ. Consider now the relative mean squared error (rMSE) between an alternative estimator $\dot{\rho}$ and the exact estimator $\hat{\rho}$ for a sample of size N, defined as the ratio (KIPNIS; DUCHI, 2017, Equation 2)

$$\mathrm{rMSE}(\dot{\rho},\hat{\rho},N) \triangleq \frac{\mathrm{MSE}(\dot{\rho},N)}{\mathrm{MSE}(\hat{\rho},N)}. \tag{4.44}$$

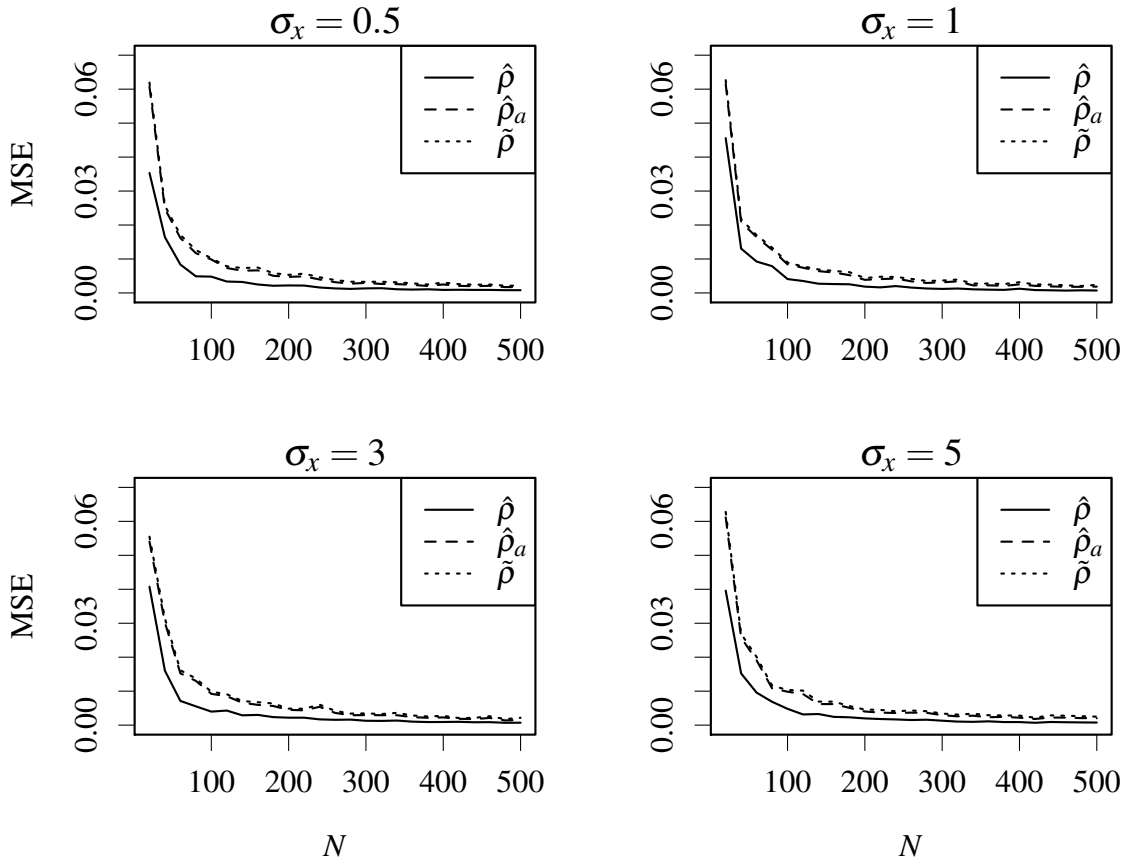

Figure 13 – MSE **of the estimators of** ρ **in** AR(1) **processes as a function of** N **for the true value of** ρ **set as** $\rho = 0.8$ **and various values of** σ_x^2 **estimated through Monte Carlo simulations.**

We used the following Monte Carlo simulation study in order to access the statistical performance of the estimators of ρ in terms of rMSE. For $\rho = 0.8$ and for each pair (N, σ_x^2), where $N \in \{20, 40, \ldots, 500\}$ and $\sigma_x \in \{0.5, 1, 3, 5\}$, we generated 200 samples of an AR(1) process with the given parameter configuration and, for each sample $r = 1, 2, \ldots, 200$, we estimated ρ with the exact, approximate and low-complexity estimators, yielding $\hat{\rho}^{(r)}$, $\hat{\rho}_a^{(r)}$, and $\tilde{\rho}^{(r)}$, for the rth sample, respectively. Then, the the MSE in (4.43) is estimated as

$$\widehat{\text{MSE}}(\hat{\rho}, N) = \frac{1}{200} \sum_{r=1}^{200} \left(\hat{\rho}^{(r)} - \rho \right)^2. \tag{4.45}$$

The rMSE in (4.44) is estimated for $\hat{\rho}_a$ and $\tilde{\rho}$ respectively as

$$\widehat{\text{rMSE}}(\hat{\rho}_a, \hat{\rho}, N) = \frac{\sum_{r=1}^{200} \left(\hat{\rho}_a^{(r)} - \rho \right)^2}{\sum_{r=1}^{200} \left(\hat{\rho}^{(r)} - \rho \right)^2},$$

$$\text{and} \quad \widehat{\text{rMSE}}(\tilde{\rho}, \hat{\rho}, N) = \frac{\sum_{r=1}^{200} \left(\tilde{\rho}^{(r)} - \rho \right)^2}{\sum_{r=1}^{200} \left(\hat{\rho}^{(r)} - \rho \right)^2}.$$

In Figure 13, we display curves of the MSE of the each estimator as a function of N. In general, the MSE of all estimators converge to zero as N grows. This empirical observation

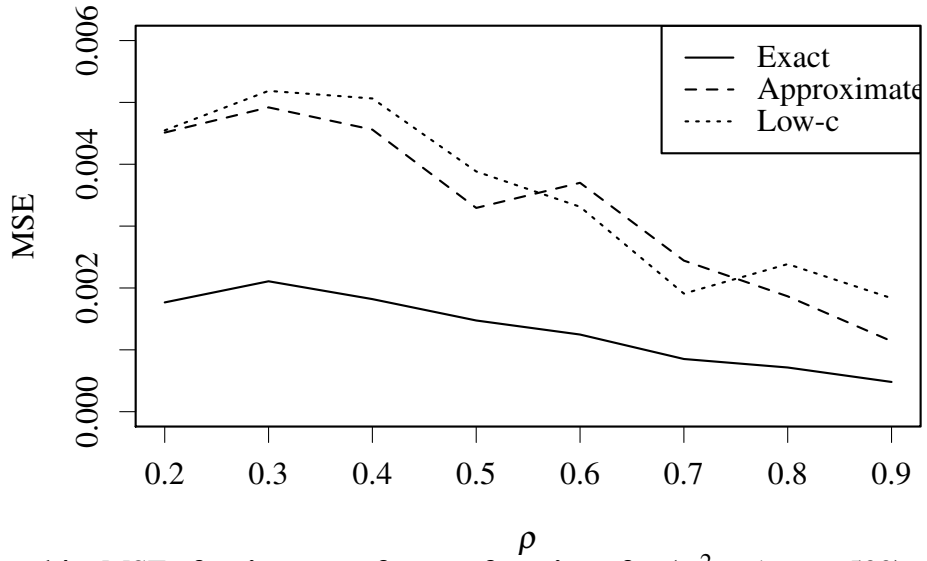

Figure 14 – MSE **of estimators of** ρ **as a function of** ρ **(**$\sigma_x^2 = 1$**,** $N = 500$**) estimated through Monte Carlo simulations.**

is in accordance with the theoretical background leading to (4.27), (4.35) and (4.36). Also, the performance of the alternative estimators is upper bounded by the performance of the exact estimator.

Qualitatively, from a visual inspection, the parameter σ_x^2 have no significant impact on the behavior of the MSE of both the exact and the approximate estimators. The MSE behavior was also homogeneous in σ_x^2 for other values of ρ. Indeed, such phenomena can be explained by the signal-to-noise ratio (SNR). The SNR is defined by the ratio between the variance, or power, of the signal y_n and the variance of the input noise x_n. In the case of AR(1) processes, the SNR is given by

$$\mathrm{SNR}_y = \frac{\sigma_y^2}{\sigma_x^2} = \frac{\sigma_x^2/(1-\rho^2)}{\sigma_x^2} = \frac{1}{1-\rho^2}. \tag{4.46}$$

We see that SNR_y does not depend on σ_x^2. Based on this observation, we set $\sigma_x^2 = 1$ for the subsequent analyses. The SNR_y metric does depend on ρ though. In fact, when $|\rho| \to 1$ SNR_y grows towards positive infinity. In Figure 14, we see that the MSE of all estimators decreases and the performances of both alternative estimators become closer to the performance of $\hat{\rho}$ as ρ approaches 1.

From a visual inspection, the curves in Figure 13 suggest that estimators $\hat{\rho}_a$ and $\tilde{\rho}$ have no significant difference in statistical behavior. In order to take a closer look into the difference amongst the estimators, we display the empirical values of the rMSE between each alternative estimator and $\hat{\rho}$ in Figure 15.

According to Figure 15, we note that in general, in the sense of MSE, the performance

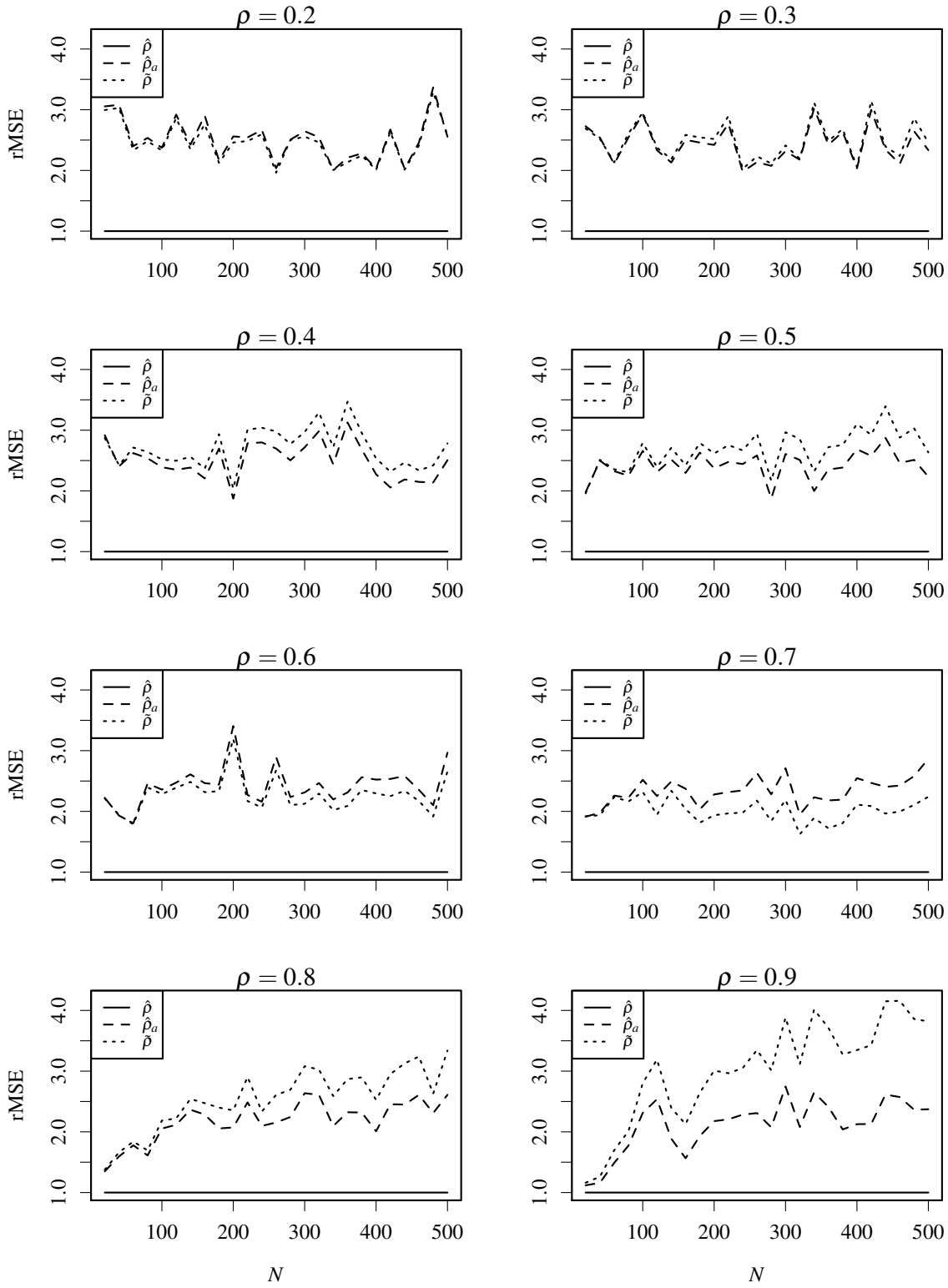

Figure 15 – rMSE **between alternative and exact estimators of** ρ **in** $\mathrm{AR}(1)$ **processes as a function of** N **for various values of** ρ **and** $\sigma_x^2 = 1$ **estimated through Monte Carlo simulations.**

of $\tilde{\rho}$ is upper bounded by the performance of $\hat{\rho}_a$, which in turn is upper bounded by the performance of $\hat{\rho}$. Particularly for $\rho \in \{0.6, 0.7\}$, the proposed low-complexity estimator $\tilde{\rho}$ behaved better than the approximate estimator $\hat{\rho}_a$. In general, the following order relation amongst the estimators of ρ can be established in terms of statistical performance as measured by the MSE:

$$\text{Exact} \succ \text{Approximate} \succeq \text{Low-complexity},$$

where $A \succ B$ means that A is preferable to B.

For $\rho = 0.9$, we note that the rMSE of $\tilde{\rho}$ grows more rapidly with N than in the other cases, which means that the difference in performance between $\tilde{\rho}$ and $\hat{\rho}$ becomes clearer as N grows (i.e., as the empirical MSE converges to its true value). That can be explained by the discontinuity in $\tilde{g}(\lambda)$ at $\lambda = 0.86$; in fact, $\tilde{g}(0.86) = 0.92$. The linear nature of the method causes the approximation having a hard time to mimic high-curvature regions, which is the case with $\lambda \to 1 \ \therefore \ \rho \to 1$. A possible solution for that problem would be to guarantee that the low-cost approximations are somewhat smooth functions.

5 APPLICATIONS

In this chapter, we experiment the developments of Chapter 4 in two applications. In the first one, we consider an extension of our developments in the theory of parameter estimation for AR(2) processes. The second application is in image processing. We consider a proof-of-concept of a simple correlation-based approach for the image segmentation problem.

5.1 LOW-COMPLEXITY INFERENCE FOR AR(2) PROCESSES

In this section, we consider the problem of estimation of the parameters of autoregressive processes of order 2, or AR(2) processes, under the low-complexity constraint. We say that y_n is an AR(2) process if it is stationary and

$$y_n = a_1 y_{n-1} + a_2 y_{n-2} + x_n \tag{5.1}$$

for some real numbers a_1, a_2, where x_n is a WN sequence with variance σ_x^2. We talk briefly about the key properties of this class of time series. In particular, we examine what is the benefit of the additional parameter a_2 in comparison to the AR(1) case, when $a_2 = 0$.

We consider the idea of iterative AR(1) filtering and how this can be used in the AR(2) parameter estimation problem. We look at the relationship between the AR(1) parameters obtained in each iteration. Specifically for AR(2) processes, we show via simulations that estimators for a_1 and a_2 can be obtained this way by applying AR(1) filtering twice. That is the main contribution of this section.

5.1.1 The PSD of AR(2) Processes

We can write Equation (5.1) as

$$a(B) \cdot y_n = \left(1 - a_1 B - a_2 B^2\right) \cdot y_n = x_n, \tag{5.2}$$

where B is the backward shift operator defined by $B^k y_n = y_{n-k}$ and $a(z) = 1 - a_1 z - a_2 z^2$ is the autoregressive polynomial (BROCKWELL; DAVIS, 2002, page 84). Let us try to invert (5.2) and write y_n as a causal filter h_k applied to x_n:

$$y_n = h(B) \cdot x_n = \sum_{k=0}^{\infty} h_k x_{n-k}, \tag{5.3}$$

where $h(z) = \sum_{k=0}^{\infty} h_k z^k$. From (5.2), since $y_n = a^{-1}(B) \cdot x_n$, we must have $h(z) = a(z)^{-1}$, or equivalently $a(z)h(z) = 1$, which yields

$$(1 - a_1 z - a_2 z^2)(h_0 + h_1 z + h_2 z^2 + h_3 z^3 + h_4 z^4 + \cdots) = 1$$
$$\therefore \quad h_0 + (h_1 - a_1)z + (h_2 - a_1 h_1 - a_2)z^2 + (h_3 - a_1 h_2 - a_2 h_1)z^3 + \cdots = 1.$$

Thus, h_k obeys the linear difference equation $h_k = a_1 h_{k-1} + a_2 h_{k-2}$, with initial conditions $h_0 = 1$ and $h_k = 0$ for $k < 0$. From this linear difference equation and the properties in (OPPENHEIM, 1999, Section 3.4), the z-transform of h_k, $H(z)$, is given by

$$H(z) = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2}}. \tag{5.4}$$

From (3.13) and (5.4), the PSD of the AR(2) process is given by

$$f_y(\omega) = \frac{\sigma_x^2}{|1 - a_1 e^{-j\omega} - a_2 e^{-j\omega}|^2} = \frac{\sigma_x^2}{1 + a_1^2 + a_2^2 - 2a_1(1 - a_2)\cos(\omega) - 2a_2\cos(2\omega)}. \tag{5.5}$$

Note that when $a_2 = 0$, $f_y(\omega)$ collapses into the AR(1) PSD (4.5). The following theorem establishes constraints on the parameters a_1 and a_2 so that (5.3) is valid, i.e., $h(B) \cdot x_n$ indeed converges to y_n.

Theorem 5.1 (Theorem 3.1.1 in (BROCKWELL; DAVIS, 2013))**.** *An unique stationary and causal solution y_n to the* AR(2) *equations* (5.1) *exists if, and only if, $a(z) \neq 0$ whenever $|z| \leq 1$.*

Following (HAMILTON, 1994, Section 2.3), we have that

$$a(z) = 0 \ \therefore \ z^{-2}a(z) = z^{-2} - a_1 z^{-1} - a_2 = 0.$$

Setting $w = z^{-1}$, assuming $z \neq 0$, we have that the requirement $|z| \geq 1$ of Theorem 5.1 for the roots of $a(z)$ is the same as requiring that the roots of $w^2 - a_1 w - a_2 = 0$, namely

$$w = \frac{a_1 \pm \sqrt{a_1^2 + 4a_2}}{2}, \tag{5.6}$$

be such that $|w| \leq 1$. Let $\Delta = a_1^2 + 4a_2$ and consider the following mutually exclusive, exhaustive cases:

1. If $\Delta = 0$, then there is only one real root with multiplicity 2 in (5.6), namely $w = a_1/2$. The condition $|w| \leq 1$ is translated into $|a_1/2| \leq 1 \ \therefore \ |a_1| \leq 2$, or, using $\Delta = 0$, since $a_1^2 = -4a_2 \leq 4$ we have the equivalent condition $a_2 \geq -1$.
2. If $\Delta > 0$, then we have has two distinct real roots in (5.6) and $|w| \leq 1$ implies that

$$a_2 \geq -1, \quad a_2 - a_1 \leq 1, \quad a_1 + a_2 \leq 1.$$

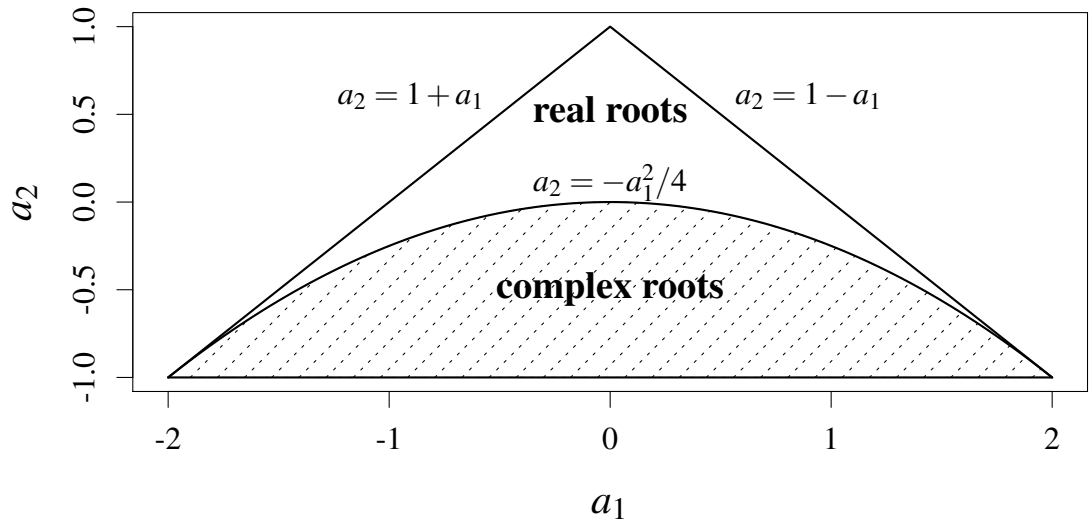

Figure 16 – The parameter space of $\mathrm{AR}(2)$ processes is known as the stability triangle due to its form in the real plane. In the area under the curve $a_2 = -a_1^2/4$, the roots (5.6) are complex numbers.

3. If $\Delta < 0$, then we have two distinct complex roots in (5.6) which are the complex conjugate of each other. Therefore, they have the same norm, namely $-a_2$, thus we must have $-a_2 \leq 1 \ \therefore \ a_2 \geq -1$.

The union of these constraints is a triangle in $\mathbb{R}^2$. It is called the stability triangle of $\mathrm{AR}(2)$ processes (HAMILTON, 1994, page 17). We display the stability triangle in Figure 16.

5.1.2 Approximate Parameter Estimation via Iterative $\mathrm{AR}(1)$ Filtering

Inspired by the technology of wavelets and filter banks (STRANG; NGUYEN, 1996; MALLAT, 2008), we consider an estimation scheme explored through Algorithm 2 as an extension of the development in the previous chapter. The similarity with wavelets is in the re-utilization of the residual series in step 3.

Algorithm 2: Algorithm used to find $\mathrm{AR}(2)$ parameter estimates based on iterative $\mathrm{AR}(1)$ filtering.

Require: $y_1, y_2, \ldots, y_N$
Ensure: Iterative $\mathrm{AR}(1)$ estimates $\hat{\rho}_1$ and $\hat{\rho}_2$

1. Fit an $\mathrm{AR}(1)$ model to the data series y_n. Let $\hat{\rho}_1$ denote the obtained estimate as in (4.25).
2. Let $\hat{x}_n = y_n - \hat{\rho}_1 y_{n-1}$, $2 \leq n \leq N$, denote the *residual series* from previous step.
3. Fit an $\mathrm{AR}(1)$ model to the data series $\hat{x}_n$. Let $\hat{\rho}_2$ denote the obtained estimate.

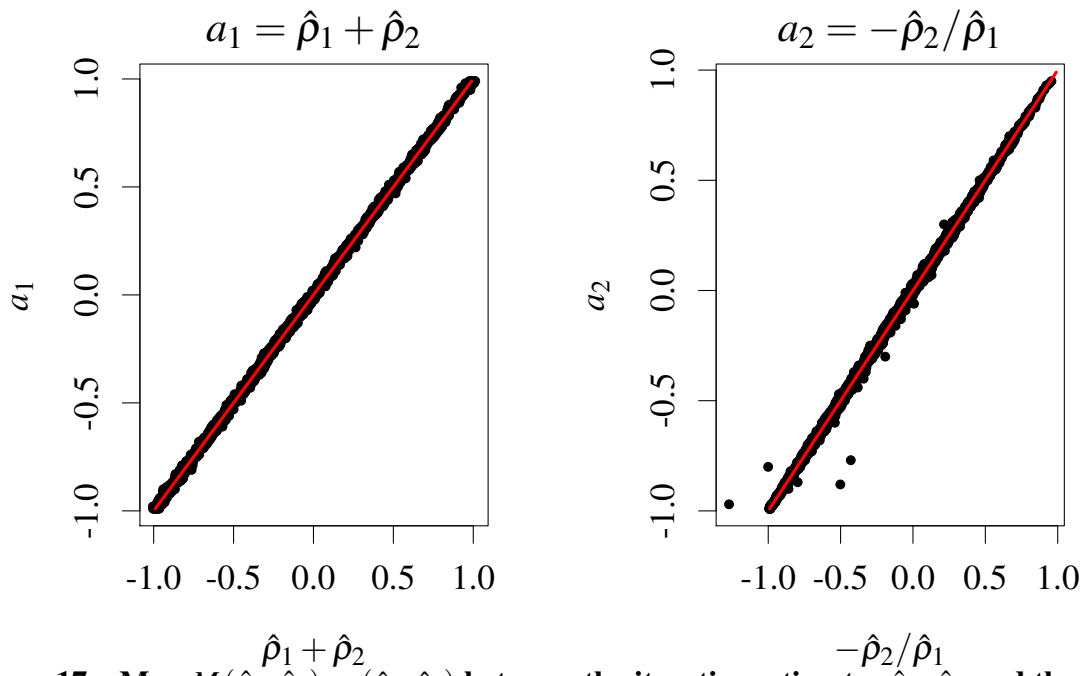

Figure 17 – Map $M(\hat{\rho}_1, \hat{\rho}_2) = (\hat{a}_1, \hat{a}_2)$ between the iterative estimates $\hat{\rho}_1$, $\hat{\rho}_2$ and the AR(2) parameters a_1 and a_2.

We argue that there is a relationship between $\hat{\rho}_1$, $\hat{\rho}_2$, a_1 and a_2. We were not able to provide a formal proof for that. We conducted a simulation study though, which suggests that such a map

$$M : (\hat{\rho}_1, \hat{\rho}_2) \mapsto (\hat{a}_1, \hat{a}_2)$$

indeed exists. The function M provides an estimator for a_1 and a_2 based on the iterative estimates $\hat{\rho}_1$ and $\hat{\rho}_2$. The goal of the Monte Carlo simulation was to discover the true relationship between these two sets of parameters. Because of that, we investigated the procedure using a large sample size, namely $N = 5000$. For each point in a grid of resolution 10^{-2} within the stability triangle in Figure 16, i.e., for each feasible point (a_1, a_2) in the AR(2) parameter space, we generated a N-point realization of an AR(2) process and obtained $(\hat{\rho}_1, \hat{\rho}_2)$ using Algorithm 2.

The results of the simulation were visually inspected using the `rgl` R package (ADLER; NENADIC; ZUCCHINI, 2003). The patterns naturally suggested that the following estimator is what we seek:

$$M(\hat{\rho}_1, \hat{\rho}_2) = (\hat{a}_1, \hat{a}_2) = \left(\hat{\rho}_1 + \hat{\rho}_2, -\frac{\hat{\rho}_2}{\hat{\rho}_1} \right). \tag{5.7}$$

In fact, in Figure 17 we display plots of $\hat{\rho}_1 + \hat{\rho}_2$ versus the true a_1 and of $-\hat{\rho}_2/\hat{\rho}_1$ versus the true a_2, along with the graphic of the functions $x \mapsto x$ and $x \mapsto -x$, respectively, in red.

Since the mapping for a_2 involves dividing $\hat{\rho}_2$ by $\hat{\rho}_1$, we may have problems of

$\Pr\{\lvert a_2-(-\hat{\rho}_2/\hat{\rho}_1)\rvert \leq q\}$	25%	50%	75%	95%	99%
Percentile q	0.004	0.008	0.014	0.025	0.036

Table 2 – Selected percentiles of the distribution of $|a_2-(-\hat{\rho}_2/\hat{\rho}_1)|$ as observed in the simulations.

numerical instability when $\hat{\rho}_1$ is close to zero. Such behavior is in fact observed in the right plot of Figure 17. We were not able to display all points in the plots of Figure 17 due to LaTeX memory issues, but, for the map $a_2 \approx -\hat{\rho}_2/\hat{\rho}_1$, errors as large as 20 are possible. We do not have a proper solution for this numerical instability issue. In Table 2, we show a selected set of percentiles of the distribution of $|a_2-(-\hat{\rho}_2/\hat{\rho}_1)|$ using the whole data from the simulations. Notice that, for instance, 99% of the absolute deviations $|a_2-(-\hat{\rho}_2/\hat{\rho}_1)|$ from the identity line are upper bounded by 0.036.

5.1.3 Performance Comparison with Maximum Likelihood Estimates

We compare the proposed iterative AR(1) filtering method with the maximum likelihood estimator (MLE) of (a_1,a_2) (BROCKWELL; DAVIS, 2013, Section 8.7). The R implementation of the MLE in the `arima` function of the `stats` package was used for the comparisons.

For each one of the feasible values of

$$(a_1,a_2) \in \{(-1.00,-0.25),(1.00,-0.50),(-0.50,0.50)\}$$

and for $N = 1024$, we simulated 500 AR(2) processes with $\sigma_x^2 = 1$. For each sample, estimates of (a_1,a_2) were obtained using (i) the MLE estimator, and the proposed iterative filtering method with (ii) the exact estimator $\hat{\rho}$ of ρ, (iii) VanVleck's approximate estimator $\hat{\rho}_a$, and (iv) the proposed low-complexity estimator $\tilde{\rho}$.

In Figures 18 and 19, we display box-plots of the biases $\hat{a}_1-a_1$ and $\hat{a}_2-a_2$ for the three selected parametric points and for each estimation method. We note that the estimates of a_2 have slightly more variability than the estimates of a_1 when the proposed iterative method is used. This phenomena is explained by the numerical instability issue of a_2 estimates discussed in the previous section. Also, using the average over the Monte Carlo replicas of the euclidian distance $\sqrt{(\hat{a}_1-a_1)^2+(\hat{a}_2-a_2)^2}$ as a figure of merit of the estimator, we can establish the

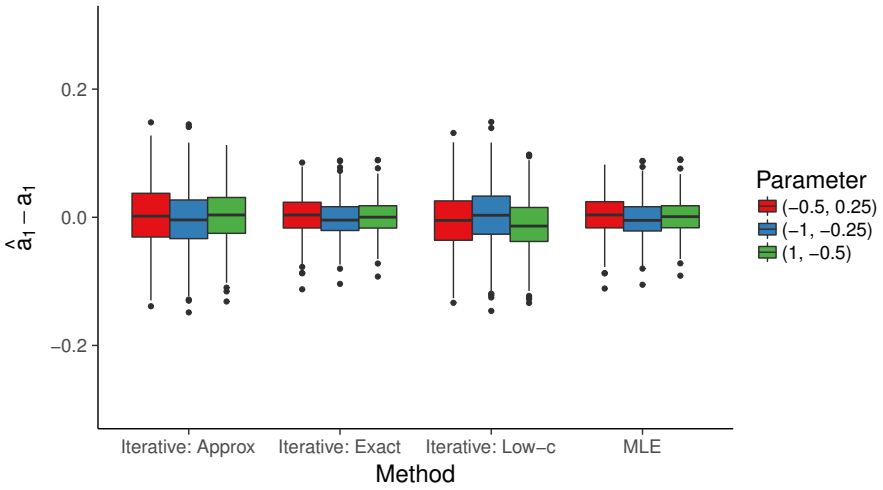

Figure 18 – Box-plots of the bias of the estimates of a_1.

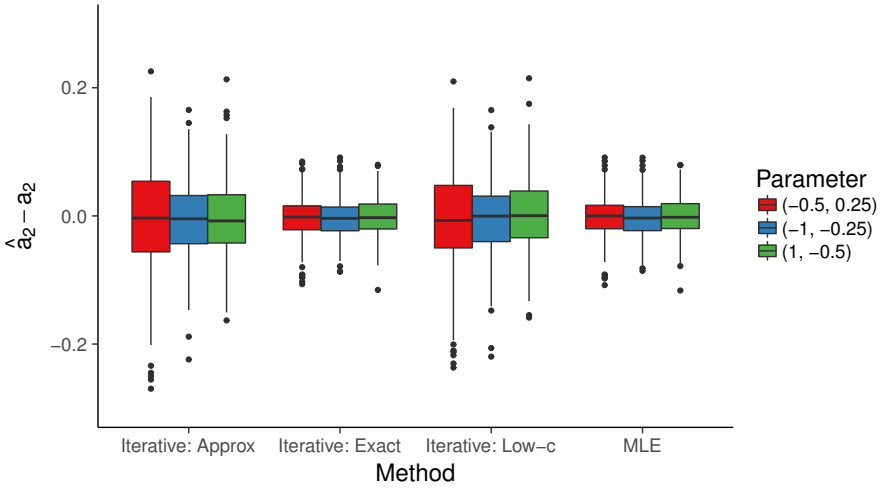

Figure 19 – Box-plots of the bias of the estimates of a_2.

following order relations:

$$(a_1, a_2) = (1.00, -0.25): \quad \text{Iterative/Exact} \succ \text{MLE} \succ \text{Iterative/Low-c} \succ \text{Iterative/Approx}$$

$$(a_1, a_2) = (-0.50, 0.25): \quad \text{MLE} \succ \text{Iterative/Exact} \succ \text{Iterative/Low-c} \succ \text{Iterative/Approx}$$

$$(a_1, a_2) = (1.00, -0.50): \quad \text{Iterative/Exact} \sim \text{MLE} \succ \text{Iterative/Low-c} \succ \text{Iterative/Approx},$$

where $A \succ B$ means that A is preferable to B and $A \sim B$ means that one is indifferent between A and B. We remark that proposed iterative scheme worked better with the proposed low-complexity estimator than with the approximate estimator in all considered cases.

5.2 IMAGE SEGMENTATION

In the image segmentation problem, we are asked for a subdivision of an input image into its constituents regions or objects (GONZALEZ; WOODS, 2007, Chapter 10). In this section, we provide a simple proof-of-concept from a correlation-based approach to the

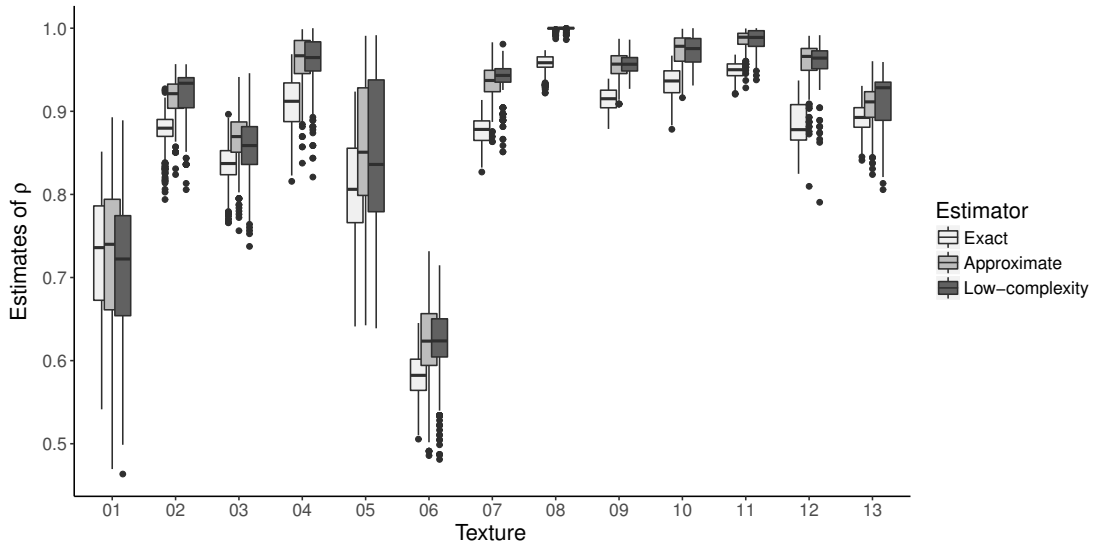

Figure 20 – Box-plots of the distribution of estimates of ρ over the rows of the 512-by-512 texture images from the USC-SIPI database.

segmentation problem. For the experiments, we used the first series of the 13 texture images of the USC-SIPI database (WEBER, 1997). The images have dimension 512-by-512 and the first series consists of unprocessed images with TIFF format.

In Figure 20, we display box-plots of the empirical distribution of estimates ρ computed for each row of the texture images from the database. We treat each row of gray pixel intensity values as the data series y_n. It seems possible to create clusters of images based solely on the information of ρ. In Table 3, we evaluate the ability of the low-complexity estimator to define clusters by comparing the variability of its estimates with the variability of the exact estimates. The variability of the estimators is measured by the sample variance ($\widehat{\text{var}(\cdot)}$) and the sample median absolute deviations from the median ($\widehat{\text{MAD}}(\cdot)$). Boldface entries indicate textures in which the proposed low-complexity estimator was more efficient in the sense of having less within-class variance and thus helping a classifier do its job. In the sense of the variance ratio, the low-complexity estimator was more efficient in 4 out of 13 cases; in the sense of the MAD ratio, the number was 6 out of 13.

In Figure 21, we display dispersion graphics of the pairs $(\rho_{\text{row}}, \rho_{\text{col}})$ of estimates of ρ over the rows and columns of the texture images. Each color represents a different texture. The "Binary" estimator uses simply $\hat{\lambda}$ with no further transformation. It is clear that clusters of textures indeed exist and are linearly separable. These observations can lead the way to fast segmentation strategies based on learning correlation intervals characterizing a given texture.

Texture	$\widehat{\text{var}}(\tilde{\rho})/\widehat{\text{var}}(\hat{\rho})$	$\widehat{\text{MAD}}(\tilde{\rho})/\widehat{\text{MAD}}(\hat{\rho})$
01	1.23	1.06
02	2.14	**0.78**
03	2.65	1.57
04	**0.79**	**0.79**
05	2.20	1.76
06	2.45	1.33
07	1.01	**0.67**
08	**0.01**	**0.00**
09	**0.75**	**0.77**
10	1.03	1.03
11	1.54	1.14
12	**0.61**	**0.47**
13	3.24	1.25

Table 3 – Relative efficiency of the proposed low-complexity estimator according to the variance (var) **and median absolute deviation** (MAD) **from the median, both computed cluster-wise. Boldface numbers indicate cases in which the low-complexity estimate had better performance.**

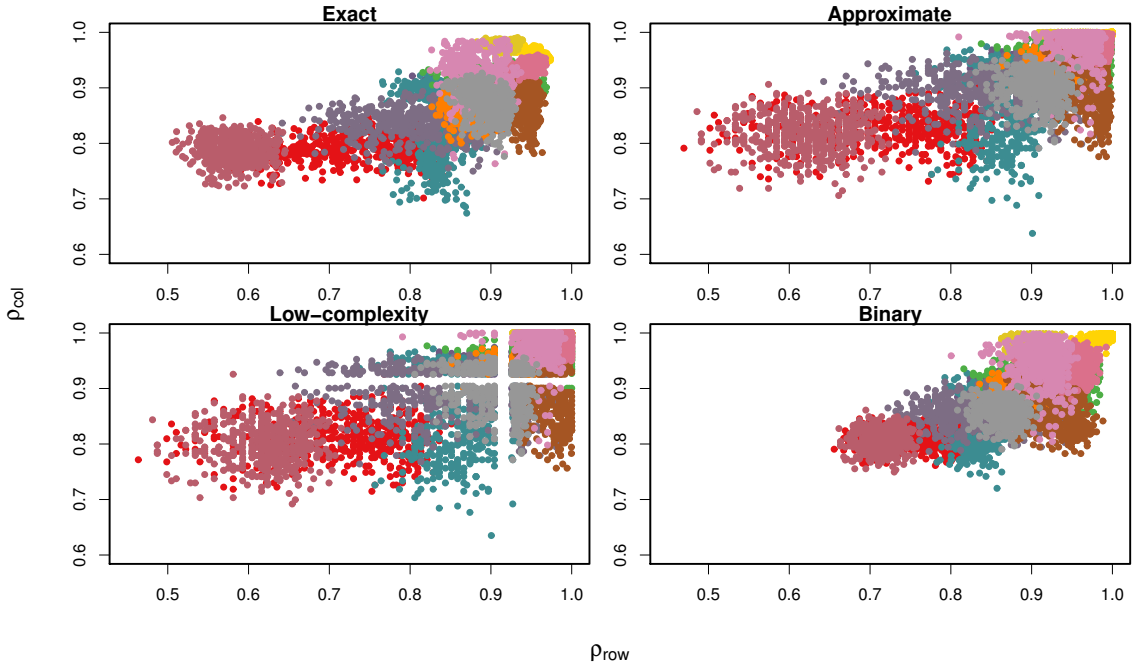

Figure 21 – Dispersion graphic of the pairs $(\rho_{\text{row}}, \rho_{\text{col}})$ of estimates of ρ over the rows and columns of the texture images of the USC-SIPI database. Each color represents a different texture.

6 CONCLUDING REMARKS

6.1 OVERVIEW OF RESULTS AND DISCUSSION

In this dissertation, we studied the parameter estimation problem in AR(1) processes under the low-complexity constraint. The work led to the following contributions:

- In Section 4.3, we studied a binarized version of AR(1) processes considering only the assumption that the process has a symmetric distribution. The symmetric scenario is less restrictive than the usual gaussian assumption. In this case, we provided the explicit formula (4.17) for the link between the transition probability λ of the binary process and the correlation structure described by ρ of the original AR(1) process. From this analysis, we conjectured that the symmetric assumption is a sufficient condition for the existence of a bijective map between λ and ρ which would allow ρ to be estimated from the one-bit observations b_n.
- In Section 4.3.2, we showed that our theoretical framework yields the same function $\rho = \cos(\pi(1-\lambda))$ already known for the case when y_n has gaussian marginals.
- In Section 4.4, we reviewed some well-known methods for the estimation of the parameters ρ and σ_x^2. In particular, by taking a closer look into the expressions for the estimators $\hat{\rho}$ and $\hat{\sigma}_x^2$, we proposed the approximate estimator $\hat{s}_x^2$ for σ_x^2 which we proved to be asymptotically equivalent to the exact estimator $\hat{\sigma}_x^2$ costing 50% less in terms arithmetical complexity.
- In Section 4.5, we proposed a low-complexity piecewise linear approximation $\tilde{g}(\lambda)$ to the curve $\rho = g(\lambda) = \cos(\pi(1-\lambda))$, which links λ and ρ in the gaussian canse. Then, a low-complexity estimator for ρ based on the one-bit observations was proposed, naturally enough, as $\tilde{\rho} = \tilde{g}(\hat{\lambda})$. Given an estimate $\hat{\lambda}$ of λ, the proposed estimator requires at most 3 sums and 6 bit-shifts. In Section 4.6, Monte Carlo simulations suggest that the proposed estimator has a statistical performance comparable to the existing alternative estimator $\hat{\rho}_a = \cos(\pi(1-\hat{\lambda}))$.
- In Section 5.1, we considered the problem of parameter estimation in AR(2) processes. Monte Carlo simulations suggest that iterative AR(1) filtering can be used to estimate the parameters of an AR(2) process. Further simulations showed that such strategy has statistical performance comparable to the MLE estimates. If we use an alternative estimator based on b_n for the iterative process, the variance of the final AR(2) estimators is slightly

larger.

- Finally, in Section 5.2, we consider a correlation-based approach to image segmentation using a database of 13 different texture images. Qualitatively, from Figure 20, it is clear that some textures can be efficiently distinguished by using only the information contained in the pixels' first autocorrelation.

6.2 FUTURE WORKS

The work can certainly be improved in many directions. We cite a few:

- The low-complexity approximation was obtained using various heuristics and optimizations via exhaustive inspection. A more rigorous optimization framework would yield better convergence guarantees. Also, the obtained curve $\tilde{g}(\lambda)$ is not smooth everywhere, as it becomes clear from Figure 21. Constraining the coefficients of B-splines (BARTELS; BEATTY; BARSKY, 1987, Chapter 4) to the set of dyadic rationals would be a more elegant way to solve the problem.
- Autoregressive models can be used to approximate the PSD of stationary signals (DJURIC *et al.*, 1999). The low-cost $\mathrm{AR}(2)$ estimator could consequently be used to provide a cheap estimator for the PSD of single tone sinusoids (SO *et al.*, 1999).

REFERENCES

ADLER, D.; NENADIC, O.; ZUCCHINI, W. RGL: A R-library for 3D visualization with OpenGL. In: **Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics, Salt Lake City**. [S.l.: s.n.], 2003. v. 35.

AHMED, N.; NATARAJAN, T.; RAO, K. R. Discrete Cosine Transform. **IEEE transactions on Computers**, IEEE, v. 100, n. 1, p. 90–93, 1974.

ALLÉVIUS, B. On the precision matrix of an irregularly sampled AR(1) process. **arXiv preprint arXiv:1801.03791**, 2018.

ANDĚL, J. Marginal distributions of autoregressive processes. In: SPRINGER. **Transactions of the Ninth Prague Conference**. [S.l.], 1983. p. 127–135.

ANDERSON, T. W.; GOODMAN, L. A. Statistical Inference about Markov Chains. **The Annals of Mathematical Statistics**, JSTOR, p. 89–110, 1957.

BARTELS, R. H.; BEATTY, J. C.; BARSKY, B. A. **An introduction to splines for use in computer graphics and geometric modeling**. [S.l.]: Morgan Kaufmann, 1987.

BASU, S.; DUCH, L.; PEÓN-QUIRÓS, M.; ATIENZA, D.; ANSALONI, G.; POZZI, L. Heterogeneous and inexact: Maximizing power efficiency of edge computing sensors for health monitoring applications. In: IEEE. **Circuits and Systems (ISCAS), 2018 IEEE International Symposium on**. [S.l.], 2018. p. 1–5.

BAYER, F. M.; CINTRA, R. J. Image compression via a fast DCT approximation. **IEEE Latin America Transactions**, IEEE, v. 8, n. 6, p. 708–713, 2010.

BETZEL, F.; KHATAMIFARD, K.; SURESH, H.; LILJA, D. J.; SARTORI, J.; KARPUZCU, U. Approximate communication: Techniques for reducing communication bottlenecks in large-scale parallel systems. **ACM Computing Surveys (CSUR)**, ACM, v. 51, n. 1, p. 1, 2018.

BILLINGSLEY, P. Statistical Methods in Markov Chains. **The Annals of Mathematical Statistics**, JSTOR, p. 12–40, 1961.

BLAHUT, R. E. **Fast algorithms for signal processing**. [S.l.]: Cambridge University Press, 2010.

BOUFOUNOS, P. T.; BARANIUK, R. G. 1-bit compressive sensing. In: IEEE. **Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on**. [S.l.], 2008. p. 16–21.

BOYCE, W. E.; DIPRIMA, R. C.; HAINES, C. W. **Elementary differential equations and boundary value problems**. [S.l.]: Wiley, New York, 2001. v. 9.

BRACEWELL, R. N. Discrete Hartley Transform. **JOSA**, Optical Society of America, v. 73, n. 12, p. 1832–1835, 1983.

BRIGGS, W. L. *et al.* **The DFT: an owners' manual for the discrete Fourier transform**. [S.l.]: SIAM, 1995.

BRITANAK, V.; YIP, P. C.; RAO, K. R. **Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations**. [S.l.]: Academic Press, 2006.

BROCKWELL, P. J.; DAVIS, R. A. **Introduction to time series and forecasting**. [S.l.]: Springer, 2002.

BROCKWELL, P. J.; DAVIS, R. A. **Time series: theory and methods**. [S.l.]: Springer Science & Business Media, 2013.

CASELLA, G.; BERGER, R. L. **Statistical Inference**. [S.l.]: Duxbury Pacific Grove, CA, 2002. v. 2.

CHVATAL, V.; CHVATAL, V. *et al.* **Linear programming**. [S.l.]: MacMillan, 1983.

CINTRA, R. J. An integer approximation method for discrete sinusoidal transforms. **Circuits, Systems, and Signal Processing**, Springer, v. 30, n. 6, p. 1481, 2011.

CINTRA, R. J.; BAYER, F. M. A DCT approximation for image compression. **IEEE Signal Processing Letters**, IEEE, v. 18, n. 10, p. 579–582, 2011.

CINTRA, R. J.; DUFFNER, S.; GARCIA, C.; LEITE, A. Low-complexity Approximate Convolutional Neural Networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, mar. 2018. Disponível em: <https://hal.archives-ouvertes.fr/hal-01727219>.

CINTRA, R. J.; OLIVEIRA, H. M. de. A short survey on arithmetic transforms and the arithmetic Hartley transform. **arXiv preprint arXiv:1504.06106**, 2015.

COOK, S. A.; AANDERAA, S. O. On the minimum computation time of functions. **Transactions of the American Mathematical Society**, JSTOR, v. 142, p. 291–314, 1969.

COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex Fourier series. **Mathematics of computation**, v. 19, n. 90, p. 297–301, 1965.

CRIBARI-NETO, F. Asymptotic inference under heteroskedasticity of unknown form. **Computational Statistics & Data Analysis, Elsevier**, Elsevier, v. 45, n. 2, p. 215–233, 2004.

DJURIC, P. M.; KAY, S. M.; VIJAY, K.; DOUGLAS, B. Spectrum estimation and modeling. **Digital Signal Processing Handbook, CRC Press LLC**, 1999.

FERGUSON, T. S. **Mathematical statistics: A decision theoretic approach**. [S.l.]: Academic Press, 2014. v. 1.

FOG, A. Instruction tables: Lists of instruction latencies, throughputs and micro-operation breakdowns for Intel, AMD and VIA CPUs. **Copenhagen University College of Engineering**, v. 93, p. 110, 2011.

FONSECA, R. V.; CRIBARI-NETO, F. Inference in a bimodal Birnbaum–Saunders model. **Mathematics and Computers in Simulation**, Elsevier, v. 146, p. 134–159, 2018.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics, New York, 2001. v. 1.

GALL, D. J. L. The MPEG video compression algorithm. **Signal Processing: Image Communication**, Elsevier, v. 4, n. 2, p. 129–140, 1992.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: CRC Press Boca Raton, FL, 2014. v. 2.

GIBBONS, J. D.; CHAKRABORTI, S. **Nonparametric Statistical Inference**. 4th. ed. [S.l.]: Marcel Dekker, 2003.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 3rd. ed. [S.l.]: Pearson Prentice Hall, 2007.

GRUNWALD, G.; HYNDMAN, R.; TEDESCO, L. A unified view of linear AR(1) models. **Research Report, Department of Statistics, University of Melbourne**, 1995.

HAMANN, B.; CHEN, J.-L. Data point selection for piecewise linear curve approximation. **Computer Aided Geometric Design**, Elsevier, v. 11, n. 3, p. 289–301, 1994.

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton University Press, Princeton, 1994. v. 2.

HAN, J.; ORSHANSKY, M. Approximate computing: An emerging paradigm for energy-efficient design. In: IEEE. **Test Symposium (ETS), 2013 18th IEEE European**. [S.l.], 2013. p. 1–6.

HAWEEL, T. I. A new square wave transform based on the DCT. **Signal processing**, Elsevier, v. 81, n. 11, p. 2309–2319, 2001.

HEIDEMAN, M. T. **Multiplicative Complexity, Convolution, and the DFT**. [S.l.]: Springer, 1988.

International Telecommunication Union. **ITU-T Recommendation H.261 Version 1: Video CODEC for Audiovisual Services at $p \times 64$ kbits**. [S.l.], 1990.

International Telecommunication Union. **ITU-T Recommendation H.263 version 1: Video Coding for Low Bit Rate Communication**. [S.l.], 1995.

JOLLIFFE, I. T. Principal component analysis for special types of data. **Principal component analysis**, Springer, p. 338–372, 2002.

JORDAN, J. Correlation algorithms, circuits and measurement applications. In: IET. **IEE Proceedings G (Electronic Circuits and Systems)**. [S.l.], 1986. v. 133, n. 1, p. 58–74.

KAY, S. M. **Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory**. [S.l.]: Prentice Hall, New Jersey, 1993.

KEDEM, B. Exact maximum likelihood estimation of the parameter in the AR(1) process after hard limiting (Corresp.). **IEEE Transactions on Information Theory**, IEEE, v. 22, n. 4, p. 491–493, 1976.

KEDEM, B. Estimation of the parameters in stationary autoregressive processes after hard limiting. **Journal of the American Statistical Association**, Taylor & Francis, v. 75, n. 369, p. 146–153, 1980.

KEDEM, B. Spectral analysis and discrimination by zero-crossings. **Proceedings of the IEEE**, IEEE, v. 74, n. 11, p. 1477–1493, 1986.

KEDEM, B.; FOKIANOS, K. **Regression models for time series analysis**. [S.l.]: Wiley Interscience, 2002.

KEDEM, B.; SLUD, E. Time series discrimination by higher order crossings. **The Annals of Statistics**, JSTOR, p. 786–794, 1982.

KEDEM, B.; YAKOWITZ, S. **Time series analysis by higher order crossings**. [S.l.]: IEEE Press, New York, 1994.

KEELEY, S.; PILLOW, J. Introduction to Gaussian Processes. 2018.

KIPNIS, A.; DUCHI, J. C. Mean estimation from adaptive one-bit measurements. **arXiv preprint arXiv:1708.00952**, 2017.

KITIC, S.; JACQUES, L.; MADHU, N.; HOPWOOD, M. P.; SPRIET, A.; VLEESCHOUWER, C. D. Consistent iterative hard thresholding for signal declipping. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on**. [S.l.], 2013. p. 5939–5943.

KLOTZ, J. Statistical inference in Bernoulli trials with dependence. **The Annals of statistics**, JSTOR, p. 373–379, 1973.

KOLMOGOROV, A. N. On logical foundations of probability theory. In: **Probability theory and mathematical statistics**. [S.l.]: Springer, 1983. p. 1–5.

KOOPMANS, L. H. **The spectral analysis of time series**. [S.l.]: Academic Press, 1995.

KULLBACK, S. **Information theory and statistics**. [S.l.]: Courier Corporation, 1997.

KWONG, J.; CHANDRAKASAN, A. P. An energy-efficient biomedical signal processing platform. **IEEE Journal of Solid-State Circuits**, IEEE, v. 46, n. 7, p. 1742–1753, 2011.

LAKSHMI, B.; DHAR, A. CORDIC architectures: A survey. **VLSI design**, Hindawi Publishing Corp., v. 2010, p. 2, 2010.

LAPSHEV, S.; HASAN, S. R. Using multiple-accumulator CMACs to improve efficiency of the X part of an input-buffered FX correlator. **Experimental Astronomy**, Springer, v. 43, n. 2, p. 177–187, 2017.

LI, T.-H.; KEDEM, B. Iterative Filtering for Multiple Frequency Estimation. **IEEE Transactions on Signal Processing**, v. 42, n. 5, 1994.

LINDSKOG, F.; MCNEIL, A.; SCHMOCK, U. Kendall's tau for elliptical distributions. In: **Credit Risk**. [S.l.]: Springer, 2003. p. 149–156.

LUTHRA, A.; SULLIVAN, G. J.; WIEGAND, T. Introduction to the special issue on the H.264/AVC video coding standard. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 13, n. 7, p. 557–559, 2003.

MACKAY, D. J.; KAY, D. J. M. **Information theory, inference and learning algorithms**. [S.l.]: Cambridge University Press, 2003.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. [S.l.]: Edusp, 2006.

MALLAT, S. **A wavelet tour of signal processing: the sparse way**. [S.l.]: Academic Press, 2008.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2012. v. 821.

MOSSBERG, M. Gaussian process parameter estimation using zero crossing data from wireless sensors. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on**. [S.l.], 2014. p. 409–413.

MOSSBERG, M.; SINN, M. Cross-correlations of zero crossings in jointly Gaussian and stationary processes with zero means. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on**. [S.l.], 2017. p. 4286–4290.

NG, E. W.; GELLER, M. A table of integrals of the error functions. **Journal of Research of the National Bureau of Standards B**, v. 73, n. 1, p. 1–20, 1969.

OPPENHEIM, A. V. **Discrete-time signal processing**. [S.l.]: Pearson Education India, 1999.

PLAN, Y.; VERSHYNIN, R. One-bit compressed sensing by linear programming. **Communications on Pure and Applied Mathematics**, Wiley Online Library, v. 66, n. 8, p. 1275–1297, 2013.

POLLARD, D. **A user's guide to measure theoretic probability**. [S.l.]: Cambridge University Press, 2002. v. 8.

PORAT, B. **Digital processing of random signals: theory and methods**. [S.l.]: Courier Dover Publications, 2008.

POURAZAD, M. T.; DOUTRE, C.; AZIMI, M.; NASIOPOULOS, P. HEVC: The new gold standard for video compression: How does HEVC compare with H.264/AVC? **IEEE consumer electronics magazine**, IEEE, v. 1, n. 3, p. 36–46, 2012.

PRATT, W. K. **Digital image processing: PIKS Scientific inside**. [S.l.]: Wiley-interscience Hoboken, New Jersey, 2007. v. 4.

PRIESTLEY, M. B. **Spectral analysis and time series**. [S.l.]: Academic Press, 1981.

RADÜNZ, A. P.; BAYER, F. M.; CINTRA, R. J. **Componentes Principais Sinalizadas com Aplicação em Processamento de Imagens**. [S.l.]: Universidade Federal de Santa Maria, Trabalho de Conclusão de Curso, 2016.

RAO, K. R.; YIP, P. **Discrete Cosine Transform: Algorithms, Advantages, Applications**. [S.l.]: Academic Press, 2014.

RÊGO, L. C. Conditioning in chaotic probabilities interpreted as a generalized Markov chain. In: **Proceedings of the 5th International Symposium on Imprecise Probability: Theories and Applications**. [S.l.: s.n.], 2007. p. 365–373.

RESCHENHOFER, E. Heteroscedasticity-robust estimation of autocorrelation. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–13, 2018.

RICE, S. O. Mathematical analysis of random noise. **Bell Labs Technical Journal**, Wiley Online Library, v. 23, n. 3, p. 282–332, 1944.

ROMNEY, J. Cross correlators. In: **Synthesis Imaging in Radio Astronomy II**. [S.l.: s.n.], 1999. v. 180, p. 57.

RUDIN, W. *et al.* **Principles of mathematical analysis**. [S.l.]: McGraw-Hill New York, 1976. v. 3.

SINN, M.; KELLER, K. Covariances of zero crossings in Gaussian processes. **Theory of Probability & Its Applications**, SIAM, v. 55, n. 3, p. 485–504, 2011.

SMITH, J. O. **Mathematics of the discrete Fourier transform (DFT): with audio applications**. [S.l.]: Julius Smith, 2007.

SO, H.; CHAN, Y.; MA, Q.; CHING, P. Comparison of various periodograms for sinusoid detection and frequency estimation. **IEEE Transactions on Aerospace and Electronic Systems**, IEEE, v. 35, n. 3, p. 945–952, 1999.

STOICA, P.; MOSES, R. L. *et al.* **Spectral analysis of signals**. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2005. v. 1.

STRANG, G.; NGUYEN, T. **Wavelets and filter banks**. [S.l.]: SIAM, 1996.

STRASSEN, V. Gaussian elimination is not optimal. **Numerische mathematik**, Springer, v. 13, n. 4, p. 354–356, 1969.

TOOM, A. L. The complexity of a scheme of functional elements realizing the multiplication of integers. In: **Soviet Mathematics Doklady**. [S.l.: s.n.], 1963. v. 3, n. 4, p. 714–716.

VLECK, J. H. V.; MIDDLETON, D. The spectrum of clipped noise. **Proceedings of the IEEE**, IEEE, v. 54, n. 1, p. 2–19, 1966.

WALLACE, G. K. The JPEG still picture compression standard. **IEEE transactions on consumer electronics**, IEEE, v. 38, n. 1, p. xviii–xxxiv, 1992.

WEBER, A. G. The USC-SIPI image database version 5. **USC-SIPI Report**, v. 315, p. 1–24, 1997.

WEISSTEIN, E. W. Prime Number Theorem. **Wolfram Research, Inc.**, 2003.

WINOGRAD, S. **Arithmetic complexity of computations**. [S.l.]: SIAM, 1980. v. 33.