A multivariate normality test for time-series, an application to sismology

Sara El Bouch

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Signal Image Parole Télécoms
Unité de recherche : Grenoble Images Parole Signal Automatique

Un test de normalité pour les séries temporelles multivariées, une application en sismologie.

A multivariate normality test for time-series, an application to sismology

Présentée par :

Sara EL BOUCH

Direction de thèse :

Olivier MICHEL — Directeur de thèse
PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes

Michel CAMPILLO — Co-directeur de thèse
Professeur, UGA

Rapporteurs :

Pierre BORGNAT
DIRECTEUR DE RECHERCHE, CNRS DELEGATION RHONE AUVERGNE

André FERRARI
PROFESSEUR DES UNIVERSITES, UNIVERSITE COTE D'AZUR

Thèse soutenue publiquement le **16 décembre 2022**, devant le jury composé de :

Pierre BORGNAT — Rapporteur
DIRECTEUR DE RECHERCHE, CNRS DELEGATION RHONE AUVERGNE

André FERRARI — Rapporteur
PROFESSEUR DES UNIVERSITES, UNIVERSITE COTE D'AZUR

Audrey GIREMUS — Examinatrice
PROFESSEUR DES UNIVERSITES, UNIVERSITE DE BORDEAUX

Romain COUILLET — Président
PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

Invités :

Laurent Deruaz-pepin
INGENIEUR, Thalès dsm

Pierre COMON
DIRECTEUR DE RECHERCHE,

MULTIVARIATE NORMALITY TEST FOR TIME-SERIES

An application to seismology

SARA EL BOUCH
Univ. Grenoble Alpes - Gipsa-lab

Sara El Bouch: *Multivariate Normality Test for Time-series, October 2022*

«Le doute est une force. Une vraie et belle force. Veille simplement qu'elle te pousse toujours en l'avant.»

Pierre Bottero, *Le Pacte des MarchOmbres, tome 2*

ABSTRACT

This thesis is concerned with time-series analysis. The present growth of interest in sensor networks and our ability to simultaneously record time series representing the fluctuations of numerous physical quantities, naturally leads to consider $\mathfrak{d}$-dimensional processes. Adapted tools for the extraction of knowledge from the ever increasing amount of recorded time-series are very much solicited. An equally exploding number of new models is developed as a response to the demand and considerable effort is put to developing efficient methods from theoretical and practical viewpoints.

Change detection is a longstanding and interdisciplinary problem at the frontier of statistics and Machine Learning (ML) practices. In particular, we are interested in the detection of rare, brief and oscillating events that appear as non-Gaussian in data recorded simultaneously on sensors. This work describes a sequential detector for non-Gaussian colored time-series embedded in Gaussian noise. To this end, we explore tools at the frontier of estimation and detection and ML practices relevant to time-series analysis.

Our major contributions consist of deriving the challenging (but relevant) limiting distribution of Mardia's Kurtosis for bivariate time-series, then extending the findings to the general multivariate case by means of random projections. The proposed results are translated to an operational sequential detector. Its performances are tested on colored copula, synthetic and real data. The good detection power of the bivariate detector is confirmed by computer experiments.

Our work is also adjacent to applications in seismology, therefore our detector is merged with this framework and applied to seismograms recorded on three-axis sensors, and arrays of sensors.

RÉSUMÉ

La présente thèse porte sur l'analyse des séries temporelles. La croissance actuelle de l'intérêt pour les réseaux de capteurs et notre capacité à enregistrer simultanément des séries temporelles représentant les fluctuations de nombreuses quantités physiques, conduit naturellement à considérer des processus d-dimensionnels. Des outils adaptés sont très sollicités pour extraire des connaissances à partir de la quantité en forte croissance des données. Un nombre tout aussi explosif de nouveaux modèles est développé pour répondre à ce besoin. Des efforts considérables sont constamment déployés pour développer des méthodes efficaces d'un point de vue théorique et pratique.

La détection des changements est un problème interdisciplinaire de longue date, à la frontière des statistiques et des pratiques de ML. Nous nous intéressons à la détection d'événements rares, brefs et oscillants qui apparaissent comme non-gaussiens dans des données enregistrées simultanément sur un réseau de capteurs. Ce travail décrit un détecteur séquentiel pour des séries temporelles colorées non gaussiennes noyées dans un bruit gaussien. À cette fin, nous explorons des outils à la frontière des méthodes de détection et estimation et des pratiques ML pertinentes pour l'analyse des séries temporelles.

Nos principales contributions consistent à dériver la difficile (mais pertinente) distribution limite du Kurtosis de Mardia pour les séries temporelles bivariées, puis à étendre les résultats au cas général multivarié au moyen de projections aléatoires bivariées. Les résultats proposés sont traduits en un détecteur séquentiel opérationnel. Ses performances sont testées sur des copules colorées, des données synthétiques et réelles. Le bon pouvoir de détection du détecteur bivarié est confirmé par des expériences numériques.

Notre travail est également ancré dans une application en sismologie, ainsi notre détecteur est fusionné avec ce cadre et appliqué aux sismogrammes enregistrés sur des capteurs à trois axes, et des réseaux de capteurs.

PUBLICATIONS

Ideas and figures have appeared previously in the following publications:

Journal publication:

1. Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *A normality test for multivariate dependent samples* in Signal Processing, Elsevier 2022

Conference proceedings:

1. Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Joint Normality test Via Two-Dimensional Projections* in ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore 2022
2. Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Multivariate Normality test for colored data* in EUSIPCO European Signal processing community, Belgrade, Serbia 2022
3. Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Un Test de Normalité pour les Processus Colorés Multivariés* in GRETSI, Nancy, France 2022

ACKNOWLEDGEMENTS

«Nous recevons sans doute plus de saluts que nous n'en donnons et l'homme s'éduque en apprenant à dire merci.»
M. Steffens, *Rien de ce qui est inhumain ne m'est étranger*

Il me faut tout d'abord remercier mes encadrants qui ont guidé mes premiers pas, sûrement maladroits, de chercheuse. Merci Pierre Comon, je me souviens de notre réunion dans la salle Chartreuse, où avec enthousiasme tu m'as montré l'expression de $\hat{\mathrm{B}}_{\mathrm{p}}$, c'est ainsi que l'aventure a commencé. Avant la plongée dans les calculs des cumulants, tu m'as mis en garde, mais je voulais absolument faire ce calcul et j'ai vraiment pris du plaisir à le faire (oui, contre toute attente). Merci Olivier Michel pour l'humain et le scientifique, merci pour nos discussions de tout et parfois rien. Merci pour ton intuition qui m'a toujours émerveillé, tes insomnies à penser mon projet de thèse et nos calculs et démonstrations de dernière minute. Merci à toi et à Christine pour votre hospitalité, pour avoir redonné le goût de la chaleur familiale à une expatriée qui s'est retrouvée seule, frontières fermées. Merci à Michel Campillo de m'inciter à écouter les chants de la terre. Cette confiance, liberté et écoute, c'est quelque chose de très précieux.
Si aujourd'hui je suis fière de mon travail et que je commence à construire mes propres opinions et directions de recherche, c'est grâce à vous.

Je remercie Pierre Borgnat et André Ferrari d'avoir accepté d'être rapporteurs de ma thèse. Vos retours et encouragements m'ont beaucoup ému et redonné confiance à l'imposteur en moi. Merci à Romain Couillet d'avoir présidé mon jury, j'espère qu'on restera en contact et qu'on échangera amplement sur le rôle des scientifiques dans la transition écologique. Je crois en notre effet papillon. Merci à Audrey Giremus d'avoir examiné mon travail et d'avoir effectué le trajet depuis Bordeaux. Énorme Merci.

La vie au labo aurait eu beaucoup moins de saveur sans les amitiés du quotidien. Merci à Cyril pour les discussions autour du repas de midi (surtout les pizzas du Canberra), les aventures à Belgrade et même mon installation à Toulouse -oui, difficile de se débarasser de

moi-. Merci au fier Dijonnais, Alex, pour les relectures, le soutien, et le kir chez Garfield. Merci à tous les doctorants qui ont partagé un bout de chemin avec moi. Merci Manon, tu m'as aidé à gérer la logistique du pot de thèse et tu as été là avec Cyril B. pour partager ma joie (et pas que) et je vous en serai reconnaissante à vie. Merci à toi, Cyril B., Rebecca et Nathalie; trouver des personnes dans des endroits improbables qui s'alignent parfaitement avec nos valeurs, c'est quelque chose de très rare et précieux.

Merci à l'équipe GAIA de m'avoir intégré -puis comme dirait Olivier m'avoir désintégré pour me laisser envoler vers d'autres horizons-. Merci pour le très bon Crumble aux pommes. Ces petites attentions changent tout. Merci à Steeve Zozor et Mathilde Radiguet; membres de mon CSI, pour votre suivi et bienveillance. Merci à Michel Desvignes et Jérôme Mars de m'avoir donné l'opportunité de faire des BEs dans mon ancienne école.

Merci à ma famille, mes parents, Khaoula, Ayoub et mes proches, aucune distance n'a jamais eu raison de votre affection et je ne vous remercierai jamais assez pour tout ce que vous avez fait pour moi.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

The next list describes several symbols that will be later used within the body of the document

$\mathbb{E}\{\}$	mathematical expectation
$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{Z}$	set of integers
δ_m^n	Kronecker delta such that $\delta_m^m = \mathbf{I}_d$ for $m = n$
$\mathbf{I}_d$	Identity matrix
$\mathbf{X}$	matrix $\mathbf{X}$ is denoted in bold font capital letters
$\odot$	Hadamard product
$\otimes$	Kronecker product
$\mathbf{0}$	vector containing only 0
$\mathbf{x}$	(column) vector in small-case bold font
$^\mathsf{T}$	transpose operator
$f_\mathbf{X}$	joint distribution of the process $\mathbf{X}$
$\mathrm{Tr}(\mathbf{X})$	trace operator of the matrix $\mathbf{X}$
x	a scalar x is denoted in small-case standard font
argmin	Argument which minimizes
Cum	cumulant
$\mathrm{diag}(\mathbf{x})$	function that takes an d-element vector $\mathbf{x}$ and outputs an d by d diagonal matrix $\mathbf{X}$ where $X_{ii} = x_i$

ACRONYMS

ML Machine Learning

LS Least Squares

RLS Recursive Least Squares

GLS Generalized Least Squares

RSS Residual Sum of Squares

IC Information Criteria

ARMA Auto-regressive Moving Average

ARIMA Auto-Regressive Integrated Moving Average

LSTM Long Short Term Memory

ANN Artificial Neural Network

RNN Recurrent Neural Network

BP Back Propagation

BPTT Back Propagation Through Time

WSS Wide Sense Stationary

SGD Stochastic Gradient Descent

RSS Residual Sum of Squares

MLE Maximum Likelihood Estimator

ESN Echo-State Network

CNN Convolutional Neural Network

MDL Minimum Description Length

GENERAL INTRODUCTION

INTRODUCTION

The work in this manuscript is rooted in a discipline: Signal processing at the border of applied mathematics, physics and computer science. This discipline has long been solicited for solving various problems from pretty much any field working with a *signal*; how to extract information from a noisy signal? How to exploit *a priori* knowledge to make the extraction more statistically significant? The answers to these questions have naturally introduced tools with an inferential flavour by attempting to construct a (parametric or non-parametric) model governing the observations. The ever increasing amount of data and the advances made in the field of computer science are causing the proliferation of unprecedented amount of models. Their aim is to skim through loads of data and extract useful information bypassing human intervention. This exercise is wrapped up under the cloak of ML. The signal processing and ML practices have the same goals and are usually combined in a favourable way to efficiently extract information from data.

New prominent paradigm

The question at the core of this manuscript is *"How to detect a low-magnitude time-series embedded in noise?"*. The difficulties of the problem to hand is the presence of large noise bursts that successfully mask the weak signals, and the absence of *a priori* information about the source generating them.

Detection and estimation

This is essentially an interdisciplinary problem to which answers have long been provided by a myriad of signal processing methods, and recently more and more ML practices. It seemed therefore natural for us to start with some pillars in signal processing and present the context in which it operated the most, governed by three properties: Linearity, stationnarity and Gaussianity, and then gradually transition to proper tools when at least one of these properties is discarded.

Before attempting to answer the question above, we add more constraints on the suggested solution: The detector must be efficient, from theoretical and computational viewpoints. It must be accompanied with guarantees on the false alarm rate. The computational burden should be low to allow the processing of large datasets.

Detector with low computational burden

The quest for satisfying the first constraint i. e. theoretical guarantees on the false alarm rate has lead us to the realm of statistics. In

order to provide statistically significant outputs, we must rigorously choose a test statistic and define its limiting distribution under a set of hypothesis. Surprisingly, there was a lack of computationally efficient procedures for testing that a multivariate time-series is Gaussian, and our quest has turned into a contribution.

Fixed false alarm rate

The main concern was and still providing an operational detector, to that end, we translate the main results of our contributions into a sequential change detector. We conclude on the performances of the detector on a set of numerical experiments. We gradually transition from synthetically generated data to real-world data.

An application to seismology

As a matter of fact, the work in this manuscript is also rooted at the border of physical applications in seismology. We focus on an application that has a long and rich history in this field: the detection of seismic tremors. The Gutenberg-Richter law [61] dictates that the cumulative number of earthquakes increases exponentially with decreasing magnitude. These events are severely drowned by seismic hum, detecting them is not an easy task but a rewarding one as it will lead to unveiling unseen patterns, and by extension to more understanding of the dynamics of Earthquakes.

In this framework, the need for large labeled datasets of past events is unmet and incompatible with detecting new waveforms; we propose to integrate the detector proposed in our work in some seismological applications. The latter will be tested on real data and compared to some earthquake detectors. However, evaluating different methods in this context becomes difficult. How to compare models on rare instances? How to compare different catalogs with a strong judgmental component?

The various facets of this work can perhaps be misleading at first. It will be a back and forth between signal processing methods, statistical hypothesis testing and probing events with ML. Then the practical applicability of the main contribution will be put to test on real-world data. The articulations of this manuscript can be resumed as follows:

> Pillars of time-series analysis (including ML practices) $\iff$ Normality test for multivariate time series $\iff$ Operational detector on synthetic and real data (exclusively on applications in seismology).

To help the reader, we detail in the following the outline and our main contributions.

OUTLINE AND CONTRIBUTIONS

Chapter 1 introduces definitions and theorems of time-series analysis to provide the reader with necessary tools to best understand

our framework and contributions. The problems in signal processing are usually solved by assuming an underlying probabilistic model. We elaborate more on this inferential flavour in the following Chapter 2. Part i serves at laying the framework of our work, and at the same time introduces the building blocks of our contributions. We were torn between presenting as many tools as possible, and focusing on the ones that will reappear in the remainder. Due to the continuous proliferation of models for different types of data, we solely focus on methods relevant to time-series, and if possible multivariate time-series. Part ii is dedicated to our original contributions, in which we propose a procedure for assessing the Gaussianity of multivariate time-series. We open it by Chapter 3 that provides a bird's eye-view on pre-existing normality tests and motivates the need to derive a novel one in our framework. For computational purposes, we focus on Higher-order statistics, more precisely on Mardia's Kurtosis and fully define the limiting distribution of this test statistic (under Gaussianity) for colored time-series. The theoretical background and the steps of the calculus are our earliest work [43]:

> Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *A normality test for multivariate dependent samples* in Signal Processing, Elsevier 2022

in which the necessary tools, theorems and main results were detailed. Chapter 4 reproduces many paragraphs and equations of [43]. The main concern of Chapter 5 is to now translate the theoretical findings to an operational real-time detector. A preliminary study first examined the generalization of our findings on bivariate processes to the general d-variate case with our second contribution: [44]

> Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Joint Normality test Via Two-Dimensional Projections* in ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, Mai 2022

in which we have conducted a comparative study between the joint normality test applied on 2D random projections and its scalar counterpart on one-dimensional projections. The results are discussed in the end of Chapter 4. Having now a practical tool to assess normality, we carry on with our interest in testing the performance of the test with or without pre-whitening; we are interested in testing the normality of regression residuals instead of raw data. Based on the findings of [46], we propose

Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Multivariate Normality test for colored data* in EUSIPCO European Signal processing community, Belgrade, Serbia, Août 2022

a two-stage operational sequential detector. Prior to applying our test, incoming data is pre-whitened using a multidimensional auto-regressive model. Numerical experiments were conducted to assess the power of the test when taking into account both the spatial and temporal dependency of the process. Encouraged by the results on synthetic data, we continued simulations on small portions of real-world data in [47]:

Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Un Test de Normalité pour les Processus Colorés Multivariés* in GRETSI, Nancy, France, Septembre 2022

These results are reported in Chapter 5. The final and third part of this work is Part **??** wherein Chapter 6 treats the merging of our works with the seismological experiments. The field is historically rich with detection methods, and recently it has attracted ML practices. It is then necessary to first review some of methods relevant to our detection task. We then carry on with simulation on real-data.

Part I

SOME PILLARS OF TIME-SERIES ANALYSIS

This first part concerns what is in one sense a small detail in the context of the vast amount of work done on *time-series analysis*. But in another sense, we are concerned with the *three dominant properties* underlying all scientific inference in signal processing: *Stationarity, linearity* and *Gaussianity*. There are a myriad of methods, depending of course, upon the end use of the analysis. As formulated eloquently in [76] *"Before passing judgement on the merits of any method, one must clearly specify for what kind of problems is a particular method intended to be used"*. Our aim in this first part is to introduce proper methods for when the aforementioned properties are verified but also, and *especially*, when one or more are *discarded*.

1

STOCHASTIC PROCESSES AND TIME-SERIES

ABSTRACT

In this first chapter, we will recall some generalities about stochastic processes. The reader will find elementary tools for the analysis of time-series both in time and spectral domains. In one sense, this odds and ends of tools and theorems constitutes the building blocks of the methods we will introduce in the remainder. More precisely, we show how Gaussianity occupies a premier place in signal processing and how it interacts with the other properties governing this realm. We introduce the Higher Order Statistics that are ubiquitous in our attempt to characterize the violation of this assumption by many real-world phenomena. In another sense, these tools naturally introduce an inferential flavour by attempting to construct a (parametric or non-parametric) model governing the observations. This point will be discussed in the second chapter under the cloak of statistical learning.

Contents

1.1 STATIONARY STOCHASTIC PROCESS

1.1.1 DEFINITIONS AND NOTATIONS

The data to hand consists of a real-valued time-series, that is, a series of available observations $x_j(n)$; $j = 1, \ldots, d$; $n = 1, \ldots, N$, made sequentially through time. Subscript j indexes the different measurements at each time point n. Although, we usually think of n as measuring the passage of time, it could also be a space variable (or even both).

We use $\mathbf{x}(n)$ for the vector with components $x_j(n)$, and is to be thought of as an observation on a real vector-valued random variable. We arrange them in:

$$\mathbf{X} \stackrel{\text{def}}{=} [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)]^T \in \mathbb{R}^{N \times d} \tag{1}$$

The *joint probability distribution* $f_{\mathbf{X}}$ of any finite subset of $\{\mathbf{x}(n)\}_{n \in \{1,\ldots,N\}}$ is prescribed. The family of all realizations together with their probability is called a *stochastic process.*

The main feature of the situations we have in mind for this work is the fact that $\forall j$, $x_j(n)$ and $x_j(m)$ will not be *independent*, for $m \neq n$.

MOMENTS AND CUMULANTS

A stochastic process is thoroughly characterized by its higher order moments. The scalar moment of order r noted $\mu^{t_1,\ldots,t_r}_{a_1,a_2,\ldots,a_r}$ reads:

$\mathbb{E}\{\prod_{k=1}^{r} (x_{a_k}(t_k) - \mu_{a_k}(t_k))\}$ *is called a central moment*

$$\mu^{t_1,\ldots,t_r}_{a_1,a_2,\ldots,a_r} = \mathbb{E}\{\prod_{k=1}^{r} x_{a_k}(t_k)\} \tag{2}$$

The first order moment vector $\boldsymbol{\mu}^n_1 = \mathbb{E}\{\mathbf{x}(n)\} \in \mathbb{R}^d$ is the mean. Similarily, the r^{th} order moment vector $\boldsymbol{\mu}^{t_1,\ldots,t_r}_r$ can be defined as the d^r vector:

$$\boldsymbol{\mu}^{t_1,\ldots,t_r}_r = [\mu^{t_1,\ldots,t_r}_{a_1,a_2,\ldots,a_r}]^T$$

where the subscripts $a_1, a_2, \ldots, a_r$ indicate the column positions of the scalar r^{th} order moments. Let $\boldsymbol{\nu} = [\nu_{a_1}(t_1), \ldots, \nu_{a_r}(t_r)]^T$ be a fixed deterministic real vector, and $\mathbf{x} = [x_{a_1}(t_1), \ldots, x_{a_r}(t_r)]^T$ a finite set of realizations of the process $\mathbf{X}$. $\Phi_{\mathbf{x}}(\boldsymbol{\nu}) = \mathbb{E}\{\exp^{j\boldsymbol{\nu}^T\mathbf{x}}\}$ is *the joint characteristic function* that uniquely specifies the joint distribution of $\mathbf{x}$. On expanding $\exp^{j\boldsymbol{\nu}^T\mathbf{x}}$ as a power series around $\boldsymbol{\nu} = \mathbf{0}$, we have:

Here $j^2 = -1$ not to be confused with an index.

$$\mu^{t_1,t_2,\ldots,t_r}_{a_1,a_2,\ldots,a_r} = (-j)^r \Big(\frac{\partial^r \Phi_{\mathbf{x}}(\boldsymbol{\nu})}{\partial \nu_{a_1}(t_1) \ldots \partial \nu_{a_r}(t_r)} \Big)_{\nu=0} \tag{3}$$

We can define $\Psi_{\boldsymbol{x}}(\boldsymbol{\nu}) = \log \Phi_{\boldsymbol{x}}(\boldsymbol{\nu})$, usually called the *(joint) second characteristic function*. It allows the definition of scalar cumulants of order r:

$$\mathrm{Cum}_{a_1,a_2,\ldots,a_r}(t_1,\ldots,t_r) = (-j)^r \Big(\frac{\partial^r \Psi_{\boldsymbol{x}}(\boldsymbol{\nu})}{\partial \nu_{a_1}(t_1)\ldots\partial \nu_{a_r}(t_r)} \Big)_{\nu=0} \quad (4)$$

When at least one index a_r is different from the rest, they are termed cross-(.) and termed auto-(.) otherwise.

The r^{th} order cumulant vector $\boldsymbol{\kappa}_r(t_1,\ldots,t_r)$ can be defined as the d^r vector:

$$\boldsymbol{\kappa}_r(t_1,\ldots,t_r) = \mathrm{Cum}(\boldsymbol{x}(t_1),\ldots,\boldsymbol{x}(t_r)) = [\mathrm{Cum}_{a_1,a_2,\ldots,a_r}(t_1,\ldots,t_r)]^{\mathsf{T}}$$

where the subscripts $a_1, a_2, \ldots, a_r$ indicate the column positions of the scalar r^{th} order cumulant. On expanding log and exp around $\boldsymbol{\nu} = \mathbf{0}$, we can identify a relationship between moments and cumulants. They are tied with the general formula of Leonov and Shiryayev [81]. Before giving some examples, we introduce some important notations.

NOTATIONS

In this work, we shall replace (2) with the less cumbersome notation:

$$\mu^{nij\ldots}_{abc\ldots} = \mathbb{E}\{x_a(n)x_b(i)x_c(j)\ldots\} \quad (5)$$

The superscripts $n, i, j\ldots$ take values in $\{1,\ldots,N\}$, and the subscripts $a, b, c, \ldots$ in $\{1,\ldots,d\}$

BRACKET NOTATION

The bracket notation was initially proposed by McCullagh [101] as a convenience to list the relationship between cumulants and moments without listing all the partitions of the indices. Perhaps, an example is more explanatory than words, the third order cumulant can be written in terms of moments as:

$$\mathrm{Cum}_{a,b,c}(n,i,j) = \mu^{nij}_{abc} - [3]\mu^{n}_{a}\mu^{ij}_{bc} + 2\mu^{n}_{a}\mu^{i}_{b}\mu^{j}_{c} \quad (6)$$

Wherein [3] $\mu^{n}_{a}\mu^{ij}_{bc} = \mu^{n}_{a}\mu^{ij}_{bc} + \mu^{i}_{b}\mu^{nj}_{ac} + \mu^{j}_{c}\mu^{ni}_{ab}$ is the sum over the three 3 partitions of three indices. If the process is zero-mean, which we will assume from now on, all partitions having a unit part $\mu^{n}_{a} = 0$ will be forced to zero and the formulae are simplified. The third-order cumulant in (6) is equal to the third order moment. However to generate the fourth-order cumulant, we need the knowledge of the fourth and second order moments i. e.

$$\mathrm{Cum}_{a,b,c,d}(n,m,i,j) = \mu^{nmij}_{abcd} - [3]\mu^{nm}_{ab}\mu^{ij}_{cd} \quad (7)$$

Important. *From now on, we will assume throughout this work that the random process* $\mathbf{X}$ *is zero-mean.*

In particular, for the second order cumulant we use the following notation:

$$\mathrm{Cum}_{a,b}(n,m) \stackrel{\text{def}}{=} S_{ab}(n,m) \tag{8}$$

for $a, b \in \{1,\dots,d\}$. We arrange the d^2 quantities of $\kappa_2(n,m)$ in a symmetric matrix, referred to as the *covariance matrix* $\mathbf{S}(n,m) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}(m)^T\}$.

$$\mathbf{S}(n,m) = \begin{pmatrix} S_{11}(n,m) & \dots & S_{1d}(n,m) \\ \vdots & \dots & \vdots \\ S_{d1}(n,m) & \dots & S_{dd}(n,m) \end{pmatrix} \tag{9}$$

1.1.2 STATIONARY PROCESSES

A special case of stochastic processes is based on the assumption that:

Definition 1.1.1 (Strict stationarity [62]) *If for all* n *and* m*, the probability distribution associated with* $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(n)$ *and* $\mathbf{x}(1+m), \mathbf{x}(2+m), \dots, \mathbf{x}(n+m)$ *remains invariant. Then the process is is said to be* strictly stationary.

This means that the statistical properties of the process are unaffected by a change of time origin, and in particular at order 2:

$$\mathbf{S}(n,m) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}(m)^T\} = \mathbf{S}(0, n-m) \tag{10}$$

We shall denote by $\mathbf{S}(\tau)$ the covariance matrix at lag $\tau = n - m$, and for $\tau = 0$, $\mathbf{S}(0) \stackrel{\text{def}}{=} \mathbf{S}$. Evidently we have:

$$\mathbf{S}(-\tau) = \mathbf{S}^T(\tau) \tag{11}$$

It is often convenient to work with the scale free (dimensionless) quantities:

$$\rho_{ab}(\tau) = \frac{S_{ab}(\tau)}{\{S_{aa}S_{bb}\}^{1/2}} \tag{12}$$

They are assembled in *the correlation matrix*

$$\boldsymbol{\rho} = \mathbf{V}^{-1/2}\mathbf{S}(\tau)\mathbf{V}^{-1/2} \tag{13}$$

where $\mathbf{V} = \mathrm{diag}([S_{11}(\tau), S_{22}(\tau), \dots, S_{dd}(\tau)]^T)$. In practice, most of statistical analysis is based solely on the second order properties and a weaker condition than strict stationarity is assumed, where only the mean function and variance $\mathbf{S}$ are supposed constant and the second order $\mathbf{S}(n,m)$ cumulant depends on the lag $|n-m|$. It's termed *weak stationarity* or Wide Sense Stationary (WSS) and customarily weak is omitted.

STATIONARY PROCESSES: ERGODIC THEORY

In theory, for a spectral density to exist and to be well defined, the dynamics generating the time-series has to be *ergodic* and allow the definition of an invariant measure.

Let $(\Omega, \mathcal{F}, P)$ denote a probability space. Let's define a measure-preserving operator T, on the space of random variables over Ω via $Tf(\omega') = f(T^{-1}\omega')$[1]. If $Tf = f$ then we say that f is invariant.

Definition 1.1.2 (Ergodicity[17]) *Let T be a measure-preserving operator of the probability space $(\Omega, \mathcal{F}, P)$. T is ergodic if and only if:*

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} P(A \cap T^{-j}B) = P(A)P(B), \quad A, B \in \mathcal{F} \tag{14}$$

Definition 1.1.3 (Ergodic process [62]) *If $\mathbf{x}(n)$ is strictly stationary with $\mathbb{E}\{|x_j(n)|\} < \infty$, $j = 1, \ldots, d$ then there is a vector, $\hat{\mathbf{x}}$, invariant such that $\mathbb{E}\{|\hat{x}_j|\} < \infty$, and*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}(j) \quad = \quad \hat{\mathbf{x}} \quad \text{a.s,} \tag{15}$$

$$\mathbb{E}\{\mathbf{x}(n)\} = \mathbb{E}\{\hat{\mathbf{x}}\} \tag{16}$$

If the condition holds, then the process $\mathbf{X}$ is said to be ergodic.

Property 1.1.1 *[62] If $\mathbf{X}$ is strictly stationary and ergodic and $\mathbb{E}\{|x_j(n)|\} < \infty$, then:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}(n) = \mathbb{E}\{\mathbf{x}(n)\} \quad \text{a.s,} \tag{17}$$

If $\mathbb{E}\{x_j(n)^2\} < \infty$ then:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{m=1}^{N} \mathbf{x}(m)\mathbf{x}(n+m)^T = \mathbf{S}(n) \quad \text{a.s,} \tag{18}$$

A weaker condition than ergodicity is the weak-mixing condition:

$$\lim_{n\to\infty} \sum_{j=1}^{n} |P(A \cap T^{-j}B) - P(A)P(B)| = 0, \quad A, B \in \mathcal{F} \tag{19}$$

A stronger condition is the *strong mixing condition* introduced by M. Rosenblatt [121]. T is strong mixing if and only $\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B)$ for $A, B \in \mathcal{F}$.

Remark. See how strong mixing implies weak mixing, and that weak-mixing implies ergodicity.

1 $\omega' \in \Omega$, the prime avoids confusion with angular frequencies defined later.

Definition 1.1.4 (strong or α-mixing process [121]) *Consider the measure of dependence* $\alpha(A,B) \overset{\text{def}}{=} \sup_{A,B\in\mathcal{F}} |P(A\cap B) - P(A)P(B)|$, *and let:*

$$\alpha_n \overset{\text{def}}{=} \sup_{i\in\mathbb{Z}} \quad \alpha(\mathcal{F}_{-\infty}^{i}, \mathcal{F}_{i+n}^{\infty}). \tag{20}$$

where $\mathcal{F}_i^j$ *denotes the* σ*-field generated by* $\mathbf{x}(k)$ *for* $i \leqslant k \leqslant j$. *The sequence* $\{\mathbf{x}(n)\}$ *is said to be strongly mixing or* α*-mixing if* $\alpha_n \to 0$ *as* $n \to \infty$.

Loosely said, we now imply that events separated in time by a lag n approach independence. This condition implies the ergodicity of the signal. Introduced as a mathematical convenience for now, it is an appealing condition from a practical point of view, since we want to believe that, let's say, seismic waves that occurred in a sufficiently distant past, are independent from the ones occuring now, and that is true for all waves.

We have introduced this condition to exhibit its link with ergodicity. We will not make use of it immediately but rather in the second part of this manuscript. Instead, we introduce a particular family of stochastic processes.

We now briefly define and recall some properties of an overwhelmingly popular class of processes: The Gaussian process that shares close links with the stationarity property introduced above.

1.1.3 GAUSSIAN PROCESSES

The grail of signal processing

A zero-mean (d-variate) random variable $\mathbf{x}$ is called Gaussian and denoted by $\mathbf{x} \sim \mathcal{N}_d(0, \mathbf{S})$ if its characteristic function has the form:

$$\Phi_{\mathbf{x}}(\boldsymbol{\nu}) = \exp\big(-\frac{1}{2}\boldsymbol{\nu}^\top \mathbf{S}^{-1}\boldsymbol{\nu}\big) \tag{21}$$

Its probability density function reads:

$$f_{\mathbf{x}} = \frac{1}{(2\pi)^{d/2}|\mathbf{S}|^{1/2}} \exp\big(-\frac{1}{2}\mathbf{x}^\top \mathbf{S}^{-1}\mathbf{x}\big) \tag{22}$$

Definition 1.1.5 (Gaussian process) *A stochastic process* $\mathbf{X}$ *is said to be Gaussian if for every finite collection* $\{i, j, \ldots, n\} \subset \{1, \ldots, N\}$, *the joint distribution of* $\{\mathbf{x}(i), \ldots, \mathbf{x}(n)\}$ *is distributed according to a multivariate Gaussian.*

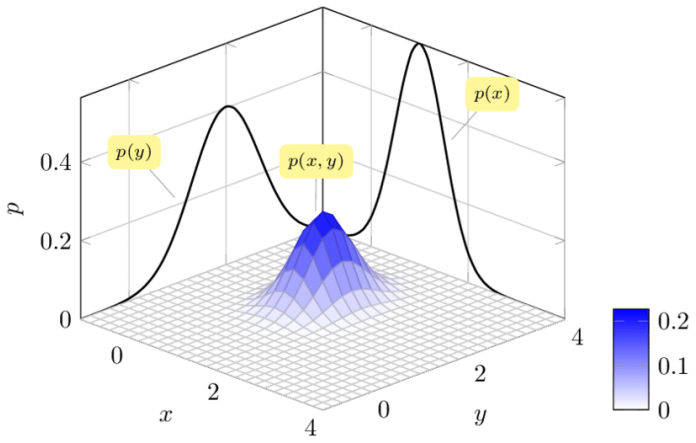


Figure 1: Illustration of the probability density function of a bivariate Gaussian variable. Image from [128].

SOME PROPERTIES[81]

- The Gaussian variable is fully described by its first and second order cumulants.
- If $\mathbf{x} \sim \mathcal{N}_{d_1}(0, \mathbf{S})$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{c} \sim \mathcal{N}_{d_2}(\mathbf{c}, \mathbf{A}\mathbf{S}\mathbf{A}^T)$. ($\mathbf{A}$ is $d_2 \times d_1$ matrix and $\mathbf{c} \in \mathbb{R}^{d_2}$). Gaussianity is stable by linear transformation.
- When approximate normality is involved, higher-order cumulants can be ignored (but not higher-order moments).
- For a zero-mean Gaussian variable, if $r = 2k + 1$, the r^{th} order moment is null. If $r = 2k$, then

$$\mu_{a_1,\ldots,a_r} = \left[\frac{(2k)!}{k!2^k}\right] \mu_{a_1 a_2} \mu_{a_3 a_4} \cdots \mu_{a_{r-1} a_r} \tag{23}$$

- As a matter of fact, we can see that that for $r = 3, 4$:

$$\mathrm{Cum}_{a_1,\ldots,a_r}(n_1, \ldots, n_r) = \mu^{n_1 \ldots n_r}_{a_1 \ldots a_r} - \mu^{n_1 \ldots n_r}_{G, a_1 \ldots a_r} \tag{24}$$

Third and Fourth order cumulants measure the extent of departure from Gaussianity.

where $\mu^{n_1 \ldots n_r}_{G, a_1 \ldots a_r}$ is the r^{th} order moment of a Gaussian signal that has the same second-order moment as $\mathbf{x}$.

- The cumulants of order $\geqslant 3$ of a Gaussian random variable are null. (as the second characteristic function of a Gaussian variable is a second order polynomial).
- Putting once again the Gaussianity in a premier place: Combining weak stationarity and Gaussianity ensures strict stationarity.

BEYOND SECOND-ORDER MOMENTS

But of course, if the process generating $\mathbf{x}(n)$ is stationary, say to the r^{th} order then:

$$\kappa_r^{t_1 \ldots t_r} = \kappa_r^{\tau_1, \tau_2, \ldots, \tau_{r-1}} \tag{25}$$

By putting $\tau_0 = 0$ and $\tau_{i-1} = t_i - t_{i-1}$ for $i \geqslant 2$. Moreover at zero-lag for $\tau_1 = \ldots = \tau_r = 0$, popular higher order statistics for scalar variables are the standardized third order and fourth order cumulants termed *Skewness* and *Kurtosis* respectively.

$$\mathcal{K}_3 = \mathbb{E}\{x(n)^3\}/(\mathbb{E}\{x(n)^2\}^{3/2} \tag{26}$$

$$\mathcal{K}_4 = \mathbb{E}\{x(n)^4\}/(\mathbb{E}\{x(n)^2\})^2 - 3 \tag{27}$$

The skewness is zero if the distribution posesses a symmetry axe. A distribution with positive $\mathcal{K}_4 > 0$ is called leptokurtic, and termed platykurtic, or platykurtotic. if $\mathcal{K}_4 < 0$.

Multivariate generalizations of Skewness and Kurtosis

Mardia [98] proposed the following measures of Skewness and Kurtosis for d-variate random variables:

$$\begin{aligned} \beta_{1,d} &= \mathbb{E}\{(\mathbf{x}_1^T \mathbf{S}^{-1} \mathbf{x}_2)^3\} \\ \beta_{2,d} &= \mathbb{E}\{(\mathbf{x}^T \mathbf{S}^{-1} \mathbf{x})^2\} \end{aligned} \tag{28}$$

Where $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent and identically distributed copies of x. We cannot assign a sign on $\sqrt{\beta_{1,d}}$, thus $\beta_{1,d}$ cannot be seen as a generalization of $\mathcal{K}_3$. Note also that $\beta_{2,d}$ is a fourth order moment and not a cumulant as $\mathcal{K}_4$.

Up to now, we have only considered the theoretical covariance functions. In practice, we have access to a finite set of observations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ from which we can obtain *consistent estimates* of the second order moments.

$$\hat{\rho}_{ab}(\tau) = \frac{\hat{S}_{ab}(\tau)}{\{\hat{S}_{aa}\hat{S}_{bb}\}^{1/2}} \tag{29}$$

$$\hat{S}_{ab}(\tau) = \frac{1}{N} \sum_{i=1}^{N-\tau} x_a(i) x_b(i+\tau) \tag{30}$$

The sample counterparts of the Skewness and Kurtosis of a multivariate population are:

$$\begin{aligned}\hat{B}_{1,d} &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{x}(i)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(j))^3 \\ \hat{B}_{2,d} &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}(n)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(n)\right)^2 \end{aligned} \tag{31}$$

1.2 LINEAR TIME INVARIANT SYSTEMS (LTI): A QUICK REVIEW

At first, we present the linear context in which the physical science has operated for most of the last two centuries. We introduce the Linear Time-invariant systems, that have the property of preserving the stationarity: If an input is stationary, the output will also inherit this property.

Each linear system is described by a function $\mathbf{H}(.)$ such that for any input *discrete-time* signal $\mathbf{x}$, the output is given by the superposition sum:

$$\mathbf{y}(n) = \sum_{\tau=-\infty}^{+\infty}\mathbf{H}(n,\tau)\mathbf{x}(\tau) \tag{32}$$

When the system is also *time-invariant*, that is $\forall n_0 \in \mathbb{Z}, \mathbf{H}(n,\tau) = \mathbf{H}(n+n_0,\tau+n_0)$, we can rewrite (32) as a discrete *convolution*:

$$\mathbf{y}(n) = \sum_{-\infty}^{+\infty}\mathbf{H}(n-\tau)\mathbf{x}(\tau) \tag{33}$$

* *denotes the convolution product* $\mathbf{y}(n) \overset{def}{=} (\mathbf{H} * \mathbf{x})(n)$

$\mathbf{H}(k)$ is a $d \times d$ matrix called the *impulse response*. The filter is said to be *causal* when $\mathbf{H}(k) = 0$ for $k < 0$ allowing the expression of the output as a function of present and past values of the input. It is said to be stable if $\sum_{k=-\infty}^{+\infty} Tr(\mathbf{H}(k)^T\mathbf{H}(k)) < \infty$ where Tr is trace operator.

The impulse response is allowed of course to be rectangular of some size $d_1 \times d_2$.

The covariance functions of the output can be thoroughly defined from that of the input along with the *impulse response* of the LTI system

$$\mathbf{S}_y(n) \overset{\text{def}}{=} \mathbb{E}\{\mathbf{y}(i)\mathbf{y}(n+i)^T\} \tag{34}$$

$$= \sum_{j=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \mathbf{H}(j)\mathbf{S}_x(n+j-k)\mathbf{H}(k)^T \tag{35}$$

$$\mathbf{S}_{xy}(n) \overset{\text{def}}{=} \mathbb{E}\{\mathbf{x}(i)\mathbf{y}(n+i)^T \tag{36}$$

$$= \sum_{j=\infty}^{\infty} \mathbf{H}(j)\mathbf{S}_x(j+n) \tag{37}$$

1.2.1 A GENERAL LINEAR PROCESS

This theorem does not imply that (38) is the true representation of the process. The latter could be non-linear or non-invertible

Theorem 1 (Wold Decomposition Theorem [62]) *Any zero-mean stationary process admits the following representation:*

$$\mathbf{x}(n) = \mathbf{u}(n) + \mathbf{v}(n) \tag{38}$$

Where

- $\mathbf{v}(n)$ *is a purely deterministic process, predictable by its own past values and* $\mathbb{E}\{\mathbf{v}(n)\boldsymbol{\epsilon}(m)^T\} = 0$. *The definition of* $\boldsymbol{\epsilon}(n)$ *will appear on the proof.*
- $\mathbf{u}(n) = \sum_0^\infty \mathbf{A}(j)\boldsymbol{\epsilon}(n-j)$ *is a weighted sum of the* $\boldsymbol{\epsilon}(n)$.

Proof. [From [62]] We introduce $\mathcal{H}$, the Hilbert space spanned by $x_j(n), j = 1, \ldots, d, n = 0, \pm 1, \ldots$, which is a real Hilbert space with inner product $\langle \mathbf{x}(n), \mathbf{y}(n) \rangle = \mathbb{E}\{\mathbf{x}(n)\mathbf{y}(n)^T\}$; we call $\mathcal{M}_n$ the closed subspace spanned by $x_j(m), m \leqslant n$. We obtain the best, in the the sense of minimizing the mean square error, h-step linear predictor of $\mathbf{x}(n+h)$ by projecting the components of that vector on $\mathcal{M}_n$. We call the error of prediction $\boldsymbol{\epsilon}(n)$. They are uncorrelated by construction i. e. we have $\mathbb{E}\{\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}(m)^T\} = \delta_m^n \boldsymbol{\Sigma}$. Choosing

$$\mathbf{A}(j) = \mathbb{E}\{\mathbf{x}(n)\boldsymbol{\epsilon}(n-j)^T\}\boldsymbol{\Sigma}^{-1}$$

such that $\sum_0^\infty \mathbf{A}(j)\boldsymbol{\Sigma}^{-1}\mathbf{A}^T(j) < \infty$. We may form

$$\mathbf{u}(n) = \sum_0^\infty \mathbf{A}(j)\boldsymbol{\epsilon}(n-j)$$

we put $\mathbf{v}(n) = \mathbf{x}(n) - \mathbf{u}(n)$, we have the previous theorem. □

Loosely said, any stationary stochastic process can be seen as the output of a causal linear filter with a white noise input. In practice, this representation is not very useful as it contains an infinite number of parameters. Thus, we introduce a special class of models.

1.2.2 SOME SPECIAL MODELS

AUTO-REGRESSIVE MOVING AVERAGE

Consider the backshift operator B such that $B^j x(i) = x(i-j)$. And consider that the matrix $A(j)$ can be approximated as $\beta(B)^{-1}\alpha(B)$ where[2] $\beta(B) = I_d - \beta_1 B - \ldots - \beta_p B^p$ and $\alpha(B) = I_d + \alpha_1 B + \ldots + \alpha_q B^q$, then we are led to consider an important class of stochastic time-series models, generated by this linear mechanism:

$$\sum_{j=0}^{p} \beta_j x(n-j) = \sum_{k=0}^{q} \alpha_k \epsilon(n-k), \quad \beta_0 = \alpha_0 = I_d, \tag{39}$$

Wherein β_j and α_k are $d \times d$ matrices. $\epsilon(n)$ satisfy:

$$\mathbb{E}\{\epsilon(n)\epsilon(n)^T\} = \Sigma \tag{40}$$

$$\mathbb{E}\{\epsilon(n)\epsilon(m)^T\} = 0 \tag{41}$$

Loosely said, $x(n)$ is determined by immediate past values of itself together with past disturbances. When $q = 0$, the process is said to be *(vector) autoregressive*, when $q > 0$ the terminology *(vector) Auto-regressive moving average* is used. When $p = 0$, the process is said to be *(vector) moving average*.

TWO SIDES OF THE SAME COIN

If all the roots of $\det \beta(B) = 0$ lie outside the unit circle, i. e.

$$\det(I_d - \beta_1 z - \ldots - \beta_p z^p) \neq 0 \quad \text{for } |z| \leqslant 1$$

then x is said to be stable and possesses the infinite causal $MA(\infty)$ representation. The stability condition ensures weak stationarity of the auto-regressive part of ARMA. If all the roots of $\det \alpha(B)$ are greater than one in absolute value then $x(n)$ is *invertible* and possesses an infinite autoregressive representation $AR(\infty)$[93].

AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE

A basic idea to extend ARMA models to non-stationary processes is presented in [22]; if the eigenvalues of the characteristic polynomial $\beta(B)$ lie on the unit circle, a variant of ARMA tackles this type of non-stationarity by an *intergration* or *differencing* operator D.

$$\beta(B)D(B)x(n) = \alpha(B)\epsilon(n) \tag{42}$$

2 $\beta(B)$ is known as the reverse characteristic polynomial of the process

where $\mathbf{D}(B) = diag([(1-B)^{k_1}, \ldots, (1-B)^{k_d}]^T)$. This model, termed *(vector) Auto-regressive Integrated Moving Average* states that after each time-series is *differenced* k_i times to reduce it to a stationary process, the resulting time-series $\mathbf{D}(B)\mathbf{x}(n)$ is an Auto-regressive Moving Average (ARMA).

ARMA processes are ubiquitous in *parametric spectral analysis* in the sense that they reformulate the problem of spectrum estimation as *What are the coefficients of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *that fit best the observations*?

For completeness, we also discuss in the following some *spectral analysis* methods.

1.3 SOME SPECTRAL TOOLS

In the history of development of time-series analysis, *time-domain methods* occupy a premier place in prediction problems. By time-domain is meant the use of autoregressive modeling, and in general means that we use initial data and not its Fourier transform. The spectral analysis methods are very appealing especially when the size of observations N is large. However, no emphasis will be put on " time domain VS. frequency domain" as these classes are complementary. Emphasis will be rather put on whether the class of proposed methods is intended for stationary or non-stationary processes.

The main aim of this section is to introduce the scattering transform that will subsequently be the basis of a learning model in Chapter 6. The definition of this transform and the understanding of its advantages require introducing the spectrogram, in particular the mel-frequency spectrogram and the wavelet transform.

1.3.1 ELEMENTARY TOOLS AND DEFINITIONS

THE (DISCRETE) FOURIER TRANSFORM

By assuming stationarity, we are confining the covariance function in a *restricted* class that has a symmetry of the group of translations of the real line, i. e. $S_{ab}(n, m) = S_{ab}(n+\tau, m+\tau)$. Now given this information about the covariance matrix, can we replace $\mathbf{x}(n)$ with a *linear* function that has a a diagonal covariance function? In some sense, the Fourier transform accomplishes just that, as $\exp(2\pi i t f)$ is an eigenfunction of the operator $U(\tau)$ which acts as: $U(\tau)f(t) = f(t+\tau)$[62].

The Discrete Time Fourier transform reads:

In practice, the finite Discrete Fourier Transform is used
$\hat{x}(\omega_k) = \sum_{n=1}^{N} x(n) \exp^{-j(n-1)\omega_k}$
with
$\omega_k = 2\pi(k-1)/N$.

$$\hat{x}(\omega) = \sum_{n=-\infty}^{+\infty} x(n) \exp^{-j\omega n} \tag{43}$$

Since the kernel of complex exponentials is separable, the Fourier transform generalizes from a one-dimensional setting to a d-dimensional one in a straightforward manner. We can view the multidimensional Fourier transform as an operator that works successively on each dimension.

THE POWER SPECTRUM

For each frequency, the squared amplitude of the Fourier transform of $\mathbf{x}(n)$ is proportional to the *power* contributed by that component and a plot of squared amplitudes against frequencies is called a *power spectrum*.

A celebrated result, known as the Wiener-Khitchine theorem states that the autocovariane function and the power spectrum are Fourier pairs, i. e.

$$\mathbf{P}_x(\omega) = \sum_{k=-\infty}^{+\infty} \mathbf{S}(k) \exp^{-j\omega k} \tag{44}$$

Spectra (or correlation) are phase-blind

for angular frequencies $\omega \in [-\pi, \pi]$. $\mathbf{P}_x(\omega)$ is a $d \times d$ matrix, it is composed of the real-values *auto-spectra* $P_{aa}, a \in \{1, \ldots, d\}$. Recall that the auto-covariance function $S_{aa}(\tau)$ is symmetric about $\tau = 0$ which means that all phase information about $x_a(k)$ is lost in $S_{aa}(\tau)$. The *cross-spectra* $P_{ab}, a, b \in \{1, \ldots, d\}, a \neq b$ on its off-diagonals are complex functions. The spectral density matrix is 2π—periodic and hermitian (complex conjugate symmetric around the zero-frequency).

The information contained in the power spectrum is essentially that which is present in the auto-correlation sequence; this would suffice for the complete statistical description of a Gaussian signal, otherwise we need to look beyond the power spectrum (auto-correlation).

HIGHER ORDER SPECTRA

Simply stated, higher-order spectra are multi-dimensional Fourier transforms of higher-order cumulants. The particular third and fourth order spectra are termed the bispectrum and the tri-spectrum.

The use of higher order spectra can be used for the suppression of Gaussian noise, to reconstruct the phase of a system and finally to characterize and detect non-linearities in the data [30, 113]. A 1-D

slice of the r^{th} order cumulant is obtained by freezing $(r-2)$ of its $r-1$ indices.

$$\mathbf{p}_r(\omega_1,\ldots,\omega_{r-1}) = \sum_{\boldsymbol{\tau}=-\infty}^{+\infty} \boldsymbol{\kappa}_r(\tau_1,\tau_2,\ldots,\tau_{r-1}) \exp^{-j(\sum_{i=1}^{r-1}\omega_i\tau_i)} \quad (45)$$

where [3] $|\omega_i| \leqslant \pi$, $\sum_{i=1}^{r-1}|\omega_i| \leqslant \pi$

PROPERTIES OF HIGHER ORDER SPECTRA

- Rigorous introduction to the r-th order spectra, symmetries and properties is given by Brillinger and Rosenblatt [23, 25].
- The previous equation (45) will be the basis of Hinich's bi-coherence measure for linearity and nonskewness (in particular Gaussianity) tests[68] detailed in Chapter 3, subsections 3.2.3 and 3.2.4.

1.3.2 SPECTRAL TOOLS FOR STATIONARY PROCESSES

MAXIMUM ENTROPY SPECTRAL ESTIMATION

The normal distribution maximizes entropy against any other distribution with the same variance

The premise of Maximum Entropy Methods (MEM) is: "When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have"[4]. The basis of this premise was stated in many intuitive forms: that distributions with higher entropy represent more "disorder" and they assume "less" according to Shannon's definition of Entropy as an information measure[76].

Consider the covariance sequence

$$S(k) = \frac{1}{N}\sum_{i=1}^{N-k} x(i)x(i+k) \quad \text{for} \quad 0 \leqslant k \leqslant p \quad (46)$$

The MEM problem is to find the probability density function f_x which has the maximum entropy subject to constraints (46) usually referred to as the *matching correlation constraints*.

Attacking this constrained optimization problem using Lagrange multipliers, see [76] for more details, the solution is of the form:

$$f_x(x(1),\ldots,x(N)) \propto \exp(-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Lambda}^{-1}\mathbf{x})) \quad (47)$$

3 $\sum_{\boldsymbol{\tau}}$ to be understood as $\sum_{\tau_1}\ldots\sum_{\tau_{r-1}}$

4 from [76]

where $\mathbf{\Lambda} = \{\lambda_{j-i}\}_{i,j\in\{1,\dots,N\}}$ is the Toeplitz matrix in which the λ_k (Lagrange multiplies) are assembled.

The Maximum Entropy distribution is thus the multivariate *Gaussian* distribution.

THE BURG MEM FORMULATION AND AR

In Burg's formulation the process is supposed to be Gaussian. Whereas the MEM constructed it.

Given a *Gaussian band-limited* process $\mathbf{x}$ and a covariance sequence $S(0),\dots,S(p)$, the problem suggested by Burg was to extend the covariance elements so that the covariance characterization would correspond to the *most random* time-series. The solution to this optimization problem yields this expression of the power spectrum [34]:

$$\hat{p}(\omega) = \frac{\sigma^2}{|\sum_{k=0}^{p} -\beta_k \exp^{j\omega k}|^2} \tag{48}$$

which is the power spectrum of an Auto-regressive process AR(p): $\sum_{k=0}^{p} \beta_k x(n-k) = \epsilon(n)$, where $\beta_0 = 1$ and $\epsilon \sim \mathcal{N}(0,\sigma^2)$. The computation of the one-dimensional maximum entropy spectrum is efficient because it can be obtained from the linear equations of autoregressive (AR) signal modeling. In the multivariate case, this relationship between maximum entropy spectrum and multi-dimensional auto-regressive modeling is no longer valid. The multidimensional maximum entropy spectrum requires the use of non-linear optimization techniques [2].

1.3.3 FOR NON-STATIONARY PROCESSES

When dealing with non-stationary signals, generally speaking the Fourier analysis is no longer valid. Consider that the phenomenon is stationary in certain intervals but its statistical properties change from one interval to the other. The obvious next step is to apply the previously presented tools on each interval separately. Now the generalization of this idea leads to the definition of a *time-varying power spectrum*. This is the idea underlying the instantaneous power spectrum proposed by Page[114], or local Fourier analysis using the *Short time Fourier transform*[51].

SPECTROGRAM

The Short time Fourier transform (STFT) reads in its univariate version:

$$\hat{x}_s(m,\omega) = \sum_{n=-\infty}^{+\infty} x(n)\phi(n-m)\exp^{-j\omega n} \tag{49}$$

Where ϕ is a window of duration T. The STFT is equivalent to shifting the input signal to zero-frequency (using $\exp^{-j\omega n}$) and applying the low-pass filter $\hat{\phi}(\omega)$.

The squared magnitude of the STFT $|\hat{x}_s(m,\omega)|^2$ yields the *spectrogram* representation of the power spectral density[24].

Naturally, arise the questions about the choice of the window, the number of time and frequency samples needed to represent $\hat{x}_s(m,\omega)$. The answer to the first is left to the practitioner, and the other two questions are answered by applying twice the Nyquist theorem. These lengths are bounded by the *uncertainty principle*.

Definition 1.3.1 (The uncertainty principle) *Given the time-spread* $\Delta_t^2 = \sum_{-\infty}^{\infty} n^2|x(n)|^2$, *and* $\Delta_\omega^2 = \sum_{-\pi}^{\pi} \omega^2\hat{x}(\omega)\hat{x}(\omega)^*$, *assuming for simplicity, but with no loss of generality, that the energy of the signal is equal to* 1, *it readily follows that:*

$$\Delta_t\Delta_\omega \geqslant \frac{1}{2}$$

Quadratic time-frequency distributions unconstrained by the uncertainty principle were proposed, such as the Wigner-Ville distribution (or its smoothed variant). This is out of the scope of this work and we refer the reader to [51].

1.3.4 BEYOND FOURIER ANALYSIS

We want to introduce the notion of frequency in a non-stationary context, we are led to seek a family of coordinates, replacing the Fourier analysis with something similar in spirit, that has an *oscillatory form* in which the notion of frequency is dominant. An example of such family is the family of *wavelets*.

WAVELET TRANSFORM

The reader familiar with wavelets will find here a brief reminder of the main definitions and notations. The reader discovering wavelets will find more details in the book of Mallat [97].

The Wavelet transform is transformation of the signal on the basis of a family of wavelets. More formally the wavelets are obtained from a *mother wavelet* by translation and hometheties:

$$\forall t \in \mathbb{R} \quad \psi_{\tau,s} = \frac{1}{\sqrt{s}}\psi(\frac{t-\tau}{s}) \tag{50}$$

where τ, s are respectively the translation and scale parameters. The mother wavelet must have finite energy and its choice fully specifies the new representation of the time-series.

An interesting interpretation of the wavelets stems from the following equations in the Fourier domain:

$$\begin{aligned}
\frac{1}{\sqrt{s}}\hat{\psi}_{s,\tau}(\omega) &= \int_{-\infty}^{+\infty} \frac{1}{s}\Psi_{s,\tau}(\frac{t-\tau}{s}) \exp^{-i\omega t}\, dt && (51)\\
&= \exp^{-i\omega\tau} \int_{-\infty}^{+\infty} \psi(t') \exp^{-i\omega t' s}\, dt' && (52)\\
&= \hat{\delta}_\tau \hat{\Psi}(s\omega) \quad , \delta_\tau = \delta(t-\tau) && (53)
\end{aligned}$$

Thus each wavelet is obtained by translating the dilated filter $\hat{\Psi}(s\omega)$ using a Dirac centered at τ. Moreover, the mother wavelet has finite energy and is therefore a band-pass filter.

In the discrete case, the scale and shift parameters are discretized as $s = s_0^k$ and $\tau = n\tau_0$ and $\Psi_{k,n}(t) = s_0^{-k/2}\psi(\frac{t-n\tau_0}{s_0^k})$ and hence $\hat{\psi}_k(\omega) = s_0^{k/2}\hat{\psi}(s_0^k\omega)$. It is customary to choose $s_0 = 2^{1/Q}$ (dyadic scale). Q is the number of wavelets per octave. Similarily to the spectrogram, the absolute value of the wavelet transform as a function of time and scale is called a *scalogram* . It also obeys the uncertainty principle. In the following paragraphs, we follow the notations of [27].

MEL-FREQUENCY SPECTROGRAM

Extensively used in speech processing, to mimic the human hearing, a mel-frequency spectrogram averages the spectrogram energy with mel-scale filterbank $\hat{\psi}_\lambda(\omega)$ obtained by dilating a complex mother wavelet with a factor $\lambda = 2^{k/Q}$:

$$Mx(m,\lambda) = \frac{1}{2\pi}\int |\hat{x}_s(m,\omega)|^2 |\hat{\psi}_\lambda(\omega)|^2 d\omega \tag{54}$$

Recall that $\hat{x}_s(t,\omega)$ is the STFT of the signal using a window ϕ of duration T. $\hat{\psi}_\lambda(\omega)$ have a constant-Q bandwidth at high frequencies [26]. Their frequency support is centered at λ with a bandwidth of the order of $\frac{\lambda}{Q}$.

Unlike the spectrogram, mel-frequency spectrogram is stable to to time-warping deformation, proof of this statement can be found in [27]. Authors in [27] have also stated that the mel-spectrogram can be approximated by time-averaging the absolute values squared of a wavelet transform, that is:

$$\begin{aligned} Mx(t,\lambda) &= \int |x(t)\phi(t-\tau) * \psi_\lambda(\nu)|^2 d\nu && (55)\\ &= \int |\int x(\tau)\phi(\tau-t)\psi_k(\nu-\tau)d\tau|^2 d\nu && (56)\\ &\approx |x * \psi_\lambda|^2 * |\phi|^2(t) && (57) \end{aligned}$$

Note that this formulation makes the time-shift invariance of the mel-spectrogram explicit. Indeed, the amount of invariance is directly controlled by the duration T of the spectrogram's window $\phi(t)$.

To reduce information loss, windows of small duration are used for averaging. However, we lose information about long-scale structures. In [27], Mallat et al. proposed a hierarchical scattering transform, that inherits the properties of the mel-spectrogram: invariance to time shifts and time-warping deformations while recovering information lost by time-averaging.

SCATTERING TRANSFORM

Let $\{\psi_\lambda(t)\} = \{\lambda\psi(\lambda t), \quad \forall k \in \mathbb{Z}, \quad \lambda = 2^{\frac{k}{Q}}\}$ be a family of wavelets obtained by dilating the mother wavelet ψ. Q is the number of filters per octave.

The zero-th order scattering coefficient denoted by S_0x is the local averaging given by $x * \phi(t)$, where $\phi(t)$ is the window of duration T. It acts as a low-pass filter thus removing the high-frequencies of the signal.

We denote by Λ_i the grid of central frequencies of $\hat{\psi}_{\lambda_i}$

The first-order scattering coefficients are obtained by convolving x with the first set of filterbanks, applying a non-linear transformation $\rho(t)$, and low-pass filtering using $\phi(t)$:

$$S_1(x(n,\lambda_1)) = \rho(x * \psi^{(1)}_{\lambda_1}) * \phi_(n), \quad \lambda_1 \in \Lambda_1 \tag{58}$$

The non-linearity $\rho(t) = |t|$, for $t \in \mathbb{C}$ is the complex modulus, and acts as a demodulation of the signal, shifting its energy to low frequencies. The other frequencies are recovered by using the second filterbank, yielding the second order coefficients:

$$S_2(x(n,\lambda_1,\lambda_2)) = \rho(\rho(x * \psi^{(1)}_{\lambda_1}) * \psi^{(2)}_{\lambda_2}) * \phi(n), \quad \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2$$

(59)

Averaging the iterated (up to order m) modulus convolution operations gives scattering coefficients of order m. However, in practice $m = 2$ is what is usually done in works, since the energy of higher-order coefficients is usually small while being computationally expensive. We thus concentrate on second order scattering representation:

$$Sx = \left(S_0 x, S_1(x(n, \lambda_1)), S_2(x(n, \lambda_1, \lambda_2))\right) \tag{60}$$

The scattering cascade of convolutions and averaging can be inter-

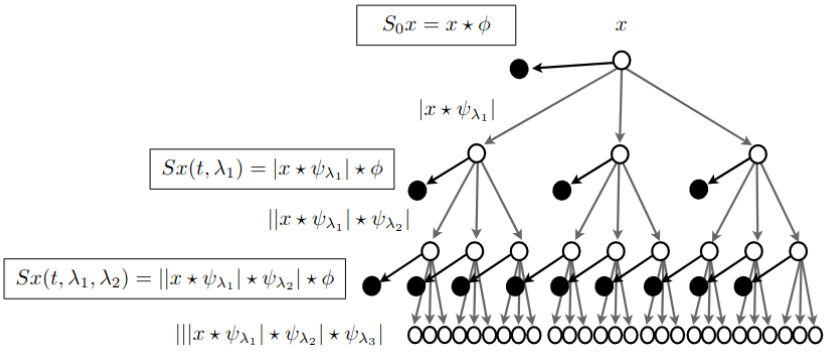


Figure 2: Illustration of scattering transform, taken from [27].

preted as a convolutional network, with important desirable properties for classification problems: stability to time shifts, time-warping deformations and additive noise [27]. It has been used in many applications [91, 129], extended to 2D applications in [132] and as we will see in the third part of this manuscript in unsupervised seismic signals classification.

1.4 CONCLUSION

To conclude, this chapter's aim is to present the necessary tools for the remainder. More precisely, we will match the presented tools here with the chapters in which they will reappear.

Higher order statistics, with a focus on Mardia's multivariate kurtosis, will be extensively exploited in Part ii. We will shed more light on the multidimensional linear auto-regressive process VAR(p) and its estimation methods in the following chapter. We will also propose other statistical methods when the assumption of linearity is discarded in Chapter 2. The spectrogram and wavelet transform will also be on the scene, either as a pre-processing tool to convert time-series to images, or as a feature-extractor using the (deep) scattering transform.

2

STATISTICAL LEARNING FOR TIME-SERIES

ABSTRACT

With the advent of computers and the information age, similar to a Cambrian explosion, statistical methods have exploded both in size and complexity. This diversification is wrapped up under the cloaks of Artificial intelligence, Machine Learning or even Data science. The Cambrian explosion metaphor is a useful one as these methods rely on the principle of trial and error. Nevertheless, it will be a risky trial and error exercise outside the realm of statistics, of estimation and detection theory. Learning is merely a specification, and estimation is rather a mechanism that actually allows it.

Contents

Due to the continuous proliferation of the learning methods and algorithms, we only focus on introducing the elements of statistical learning that will be used in the remainder. We focus on the problem of time-series forecasting because the data to hand consists of multivariate time-series (we choose to work in the time-domain), and the prediction problem will be later on related to our primary task i. e. detection of signals. We choose to elaborate on the Vector Autoregressive model VAR(p) and its estimation methods as it is a building block of the sequential detector we introduce in Chapter 5 and apply later on real data. Aware that the linearity assumption may be too simplistic for complex real-data, we present some non-linear statistical models. There is a great deal of *hype* surrounding Artificial Neural Network (ANN), this terminology encompasses a myriad of architectures that go beyond the scope of this work, here we review rapidly some relevant theory to ANN and describe the architectures relevant to time-series analysis.

2.1 BASIC IDEAS IN MODEL BUILDING

If an observed phenomenon is completely understood, it might be possible to derive mathematical expressions that thoroughly describe it. However, when the complete knowledge is not available or incomplete, we need to resort to a *an empirical model*. These two extremes of pure theory and practice interact; and a model from a class of convenient mathematical functions is considered.

Recently, we are drowning in information that comes in the form of increasing amounts of data; and models are solicited more than ever to extract knowledge. This has led to an explosion of statistical models, and the new developments were brought up by researchers from various fields: computer science, engineering, statistics and signal processing with one common exercise: learning from data. The main purpose of this chapter is to explain learning ideas from ML in a statistical framework. One should distinguish between supervised and unsupervised learning. In this work, we focus on the supervised paradigm summarized as:

- We use incomplete knowledge about the phenomenon to postulate a *suitable class of models*, and choose one of them. For example, suppose this model is reasonable $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}$.
- The model is then *fitted* to data and the parameters are *estimated*. We attempt to learn f from a training set $\{x_i, y_i\}$ fed to a learning algorithm.

Fitting means that the number and numerical values of the parameters will be estimated from observations

- The *adequacy* of the model is put to test to uncover a possible lack of fit.
- Upon completion of the learning process, if the model is judged adequate, it is used to *infer* knowledge on the rest of the available data i. e. in response to a new input x_i, the learning algorithm outputs an estimation $\hat{f}(x_i)$.

This exercise is called *supervised learning*. The approach taken in statistics has been from the perspective of estimation theory and function approximation. Thus, we start this chapter by laying down the estimation framework, fundamental to define a criteria of learning and assessing the best model according to some metric that matches the observations.

2.1.1 THE MATHEMATICAL ESTIMATION PROBLEM

The first step is to describe the stochastic process by its probability density function $f_X(\beta)$. The PDF is parameterized, thus affected, by the unknown parameter β. The goal of estimation is to *infer* the value of β from the set of observations. The *estimator* may be thought of as a rule that assigns, for each realization of $\mathbf{X}$, a value to β. This *estimator* is itself a random variable and thus its performance can only be described statistically.

In search of an optimal estimator, one must define a criterion. Usually by minimizing the mean square error (MSE):

$$\mathrm{MSE}(\hat{\beta}) \stackrel{\text{def}}{=} \mathbb{E}\{(\hat{\beta}-\beta)^2\} = \mathbb{V}\mathrm{ar}(\hat{\beta}) + \underbrace{(\mathbb{E}\{\hat{\beta}\}-\beta)^2}_{b(\hat{\beta})} \tag{61}$$

Bias-variance trade-off

From a practical viewpoint, because of the trade-off between the first and second-hand term (referred to as the *bias* $b(\hat{\beta})$), one often searches for a realizable estimator on the restricted class of *unbiased* estimators, which minimizes $\mathbb{V}\mathrm{ar}(\hat{\beta})$, termed the Minimum Variance Unbiased Estimator (MVU).

There is no turn-the-crank procedure to find the MVU, but several approaches were proposed, in which the names of Cramér-Rao, and Rao-Blackwell occupy a premier place [80]. Another procedure is to restrict the class of estimators by using a mathematical convenience, for example *linearity*.

2.1.2 LINEAR MODEL

The linear model has been a mainstay of statistics over the last two decades. Given a vector of inputs $\mathbf{x} = [x_1, \ldots, x_p]^T$, we predict the output y via the model:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{62}$$

The term $\hat{\beta}_0$ is the intercept, also known as the *bias* in ML terminology. The linear model can be written in vector format as a inner product between $\mathbf{x} = [1, x_1, \ldots, x_p]^T$ and $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]$:

$$y = \mathbf{x}^T \boldsymbol{\beta} \tag{63}$$

In general, to model d outputs, y can be a d-dimensional vector, in which case β would a be a $p \times d$ matrix.

Given a set of training data of size N that is a set of measurements $\{(\mathbf{x}_i, y_i)\}_{i \in 1,\ldots,N}$. The most popular approach is to pick the coefficients β that minimize the Residual Sum of Squares (RSS).

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2 \tag{64}$$

The solution is easiest to characterize in matrix notation:

$$\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \tag{65}$$

where $\mathbf{Z}$ is an $N \times p$ matrix with each row an input vector $\mathbf{x}_i \in \mathbb{R}^p$. Differentiating w.r.t $\boldsymbol{\beta}$, we get the normal equations:

$$\mathbf{Z}^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) = 0 \tag{66}$$

If $\boldsymbol{\Gamma} = \mathbf{Z}^T \mathbf{Z}$ is non-singular, then the unique solution is given by:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^{-1} \mathbf{Z}^T \mathbf{y} \tag{67}$$

MAXIMUM LIKELIHOOD ESTIMATION

A more general principle for estimation than minimizing the residual sum of squares, is the maximum likelihood estimation principle. This intuitive method is overwhelmingly the most popular approach to obtain practical estimators. The Maximum Likelihood Estimator (MLE) is defined as the parameter for which the observed data have the highest joint probability for $\mathbf{x}$ fixed:

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{x}) = f_{\mathbf{x}}(\boldsymbol{\beta})$$

.

If we make two assumptions on the linear problem with additive error $y = f(x) + \epsilon$: (i) the observations are independent and drawn from the same probability distribution (i.i.d.) and (ii) the target variable y has statistical noise with a Gaussian distribution i. e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$; then it is equivalent to maximum likelihood estimation using the negative log-likelihood expressed as:

$$-\log \mathcal{L}(\beta, x) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i{}^T\beta)^2 \quad (68)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} -\log \mathcal{L}(\beta, x). \quad (69)$$

The last term involving β is RSS(β) defined in Equation 64 up to a multiplier. Differentiating (68) w.r.t β and equating the result to 0 yields the normal equation $Z^T(y - Z\beta) = 0$. and the maximum likelihood estimator $\hat{\beta}$ is similar to the one obtained by Least Squares (LS) for the additive error model in Equation 67.

As stated in Theorem 7.1 in [80], the MLE $\hat{\beta}$ is asymptotically (for large enough data) optimal: It is unbiased, achieves the Cramér-Rao lower bound and have a Gaussian probability density function.

The outputs y_i vary in nature. This distinction has led to a naming convention:

- Regression when we predict quantitative outputs ($y_i \in \mathbb{R}$).
- Classification when we predict qualitative outputs (y_i discrete).

Let's first look at an example of the linear model in a regression problem for time-series, when this dimension of time is added, this task is often referred to as *time-series prediction* or *forecasting*.

2.2 TIME-SERIES FORECASTING

Let $x(i) \in \mathbb{R}^d$ be a random realization of the process X at time i. Formally, the approach to forecasting time-series, up to a *horizon* h, is expressed as follows

$$\hat{x}(t+h) = f(x(t), x(t-1), \ldots) \quad (70)$$

Naturally arise two questions about the choice of $f(.)$ and how many past observations should be accounted for in the forecast. We first start by restricting $f(.)$ to the class of linear functions, namely, the Vector Autoregressive model VAR(p). We relax this assumption using the Kernel trick as proposed by [77]. Finally, we introduce some neural network architectures relevant to time-series forecasting.

2.2.1 LINEAR APPROACH: FOCUS ON VAR(p)

There exists a large literature on linear models, in which the Vector Auto-regressive Model VAR(p) is one of the most successful and easy to use model for the analysis of multivariate time-series. In the following, we extend the discussion on paramerter estimation of VAR(p) model because it will be used extensively in numerical simulations in Chapter 5,Chapter 6.

The VAR(p) model is a simple case of the vector ARMA models presented earlier in subsection 1.2.2 in Chapter 1. Recall the Wold Decomposition theorem stated in Theorem 1 which ensures that under general conditions, any system can be expressed as an MA(∞) process (if we assume that the only deterministic component in Equation 38 of the system is the mean). We have also seen that MA(∞) and AR(∞) are two sides of the same coin, and that under invertibility conditions, MA(∞) can be expressed as a VAR(∞) which can be expressed well by a finite VAR(p). This result is powerful and demonstrates the generality of the processes under study.

We recall here the definition of a VAR(p) process:

$$\mathbf{x}(i) = \mathbf{b} + \sum_{i=1}^{p} \boldsymbol{\beta}_i \mathbf{x}(n-i) + \boldsymbol{\epsilon}(i) \tag{71}$$

where $\mathbf{x}(i)$ a d-dimensional random variable, $\boldsymbol{\beta}_i$ are fixed $d \times d$ matrices, $\boldsymbol{\epsilon}(i)$ is a realization of a WSS process, whose second order cumulant is denote $\boldsymbol{\Sigma}$. Finally, $\mathbf{b}$ is the intercept.

MULTIVARIATE LEAST SQUARES

Assuming for now that the number of past values p is known, finding the coefficients $\boldsymbol{\beta}_i$ that minimize RSS is easiest using the matrix form:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{72}$$

or using the column stacking operator vec:

$$\begin{aligned} vec(\mathbf{Y}^T) &= vec(\boldsymbol{\beta}^T \mathbf{Z}^T) + vec(\boldsymbol{\epsilon}^T), & (73)\\ &= (\mathbf{Z}^T \otimes \mathbf{I}_d) vec(\boldsymbol{\beta}^T) + vec(\boldsymbol{\epsilon}^T) & (74) \end{aligned}$$

where, $\forall i \in \{p, \dots, N-1\}$:

$$\begin{aligned}
\boldsymbol{\beta} &= \left[\mathbf{b}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_p\right]^T \in \mathbb{R}^{(dp+1)\times d} \\
\mathbf{Y} &= \left[\mathbf{x}(p+1), \dots, \mathbf{x}(N)\right]^T \in \mathbb{R}^{(N-p)\times d} \\
\mathbf{Z} &= \left[\mathbf{z}(p), \mathbf{z}(p+1), \dots, \mathbf{z}(N-1)\right]^T \in \mathbb{R}^{(N-p)\times(dp+1)} \\
\mathbf{z}(i) &= \left[1; \mathbf{x}(i); \mathbf{x}(i-1); \dots; \mathbf{x}(i-p+1)\right] \in \mathbb{R}^{(dp+1)\times 1} \\
\boldsymbol{\epsilon} &= \left[\boldsymbol{\epsilon}(p+1), \boldsymbol{\epsilon}(p+2), \dots, \boldsymbol{\epsilon}(N)\right]^T \in \mathbb{R}^{(N-p)\times d}
\end{aligned}$$

Recall that $\boldsymbol{\Sigma} = \mathbb{E}\{\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}(n)^T\}$. We can perform a Generalized Least Squares (GLS) to obtain the estimator $\hat{\boldsymbol{\beta}}$ that minimizes the cost function $J(\boldsymbol{\beta})$:

$$J(\boldsymbol{\beta}) = \mathrm{Tr}\left((\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^T\right) \tag{75}$$

Hence by differentiation with respect to $\boldsymbol{\beta}$ and equating to zero, we obtain the estimator [94]:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^{-1}\mathbf{Z}^T\mathbf{Y} \tag{76}$$

assuming of course that $\boldsymbol{\Gamma} = \mathbf{Z}^T\mathbf{Z}$ is non-singular.

It is worth noting that the GLS is equivalent to minimizing the RSS on each equation of the following linear system:

$$\begin{cases}
y_1(n) = \beta^1_{1,1}x_1(n-1) + \beta^1_{1,2}x_2(n-1) + \dots + +\beta^p_{1,d}x_d(n-p) + \epsilon_1(n) \\
y_2(n) = \beta^1_{2,1}x_1(n-1) + a\beta^1_{2,2}x_2(n-1) + \dots + \beta^p_{2,d}x_d(n-p) + \epsilon_2(n) \\
\quad\vdots \\
y_d(n) = \beta^1_{d,1}x_1(n-1) + \beta^1_{d,2}x_2(n-1) + \dots + \beta^p_{d,d}x_d(n-p) + \epsilon_d(n)
\end{cases}$$

where $\beta^k_{i,j}$ denotes the scalar coefficients of the d by d parameter matrix $\boldsymbol{\beta}_k$.

This result is due to [144] who showed that GLS and LS estimation in a multiple equation model are identical if the regressors in all equations are the same.

RECURSIVE LEAST SQUARES

To allow the model to track potential non-stationarities in the observations, while keeping the computational complexity (resulting from the inversion of the matrix $\boldsymbol{\Gamma}$ at each time-step) low, we recall here the well-known Recursive Least Squares algorithm: Let $\hat{\boldsymbol{\beta}}_k$ be the k^{th} column of $\widehat{\boldsymbol{\beta}}$ defined in (76), and $\mathbf{y}_k(t-1) = \left[x_k(1), x_k(2), \dots, x_k(t-1)\right]^T$.

$$\mathbf{y}_k(t-1) = \mathbf{Z}\hat{\boldsymbol{\beta}}_k(t-1) + \boldsymbol{\epsilon}_k(t-1) \tag{77}$$

$$\hat{\boldsymbol{\beta}}_k(t-1) = \boldsymbol{\Gamma}^{-1}(t-1)\mathbf{Z}^T\mathbf{y}_k(t-1) \tag{78}$$

Suppose we want to update the model with new observations $\mathbf{x}(t)$; a new row $\mathbf{z}(t)^T$ is appended to $\mathbf{Z}$ in (75), and a new observation $x_k(t)$ is appended to $\mathbf{y}_k$. In the Recursive Least Squares (RLS) algorithm, $\mathbf{\Gamma}^{-1}(t)$ and $\hat{\boldsymbol{\beta}}_k(t)$ are recursively expressed for $t > N$ as:

$$\begin{aligned} \mathbf{\Gamma}^{-1}(t) &= \left(\lambda_1 \mathbf{\Gamma}(t-1) + \mathbf{z}(t)\mathbf{z}(t)^T\right)^{-1} && (79)\\ &= \lambda_1^{-1}\mathbf{\Gamma}^{-1}(t-1) - b\mathbf{u}\mathbf{u}^T && (80)\end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{w}}_k(t) &= \mathbf{\Gamma}^{-1}(t)\tilde{\mathbf{Z}}^T\tilde{\mathbf{y}}_k && (81)\\ &= \hat{\boldsymbol{w}}_k(t-1) - b(\mathbf{z}^T(\hat{\boldsymbol{w}}_k(t-1) + x_k(t)\mathbf{u}))\mathbf{u} + x_k(t)\mathbf{u}, \end{aligned}$$

$\mathbf{u} = \lambda_1^{-1}\mathbf{\Gamma}^{-1}(t-1)\mathbf{z}(t)$ and $b = (1 + \mathbf{z}(t)^T\mathbf{u})^{-1}$. Note that $\tilde{\mathbf{Z}}$ is $\mathbf{Z}$ in (75) augmented with the row $\mathbf{z}(t)^T$, and $\tilde{\mathbf{y}}_k = \left[\mathbf{y}_k(t-1); x_k(t)\right]$. The fading factor λ_1 allows more flexibility in the model by discounting exponentially past information. Other variants allow more flexibility by updating the value of the fading factor and the order p of the model, see e. g. [57], [28].

PROPERTIES OF ESTIMATOR

Proposition 2.2.1 (Asymptotic properties of the LS estimator[93]) *Let $\mathbf{x}(n)$ be a stable, stationary VAR(p) process with standard WSS noise, and $\hat{\boldsymbol{\beta}} = \mathbf{\Gamma}^{-1}\mathbf{Z}^T\mathbf{Y}$ is the LS estimator of the parameters $\boldsymbol{\beta}$. Then $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Additionally:*

$$\begin{aligned} \operatorname*{plim}_{N\to\infty}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} && (82)\\ \sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\to \mathcal{N}(0, \mathbf{I}_d \otimes \mathbf{\Gamma}^{-1}) && (83)\end{aligned}$$

Here plim denotes convergence in probability and $\mathbf{\Gamma} \overset{\text{def}}{=} \operatorname*{plim}_{N\to\infty}\mathbf{Z}^T\mathbf{Z}/N$ is supposed non-singular.

If $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{\Sigma})$, consistency and asymptotic normality of the LS estimator are ensured for Gaussian (stable) VAR(p) processes [94].

In practice, $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are estimated. An obvious consistent estimator of $\mathbf{\Gamma}$ is $\hat{\mathbf{\Gamma}} = \frac{\mathbf{Z}^T\mathbf{Z}}{N}$ and a consistent estimator of the covariance matrix of the white noise is:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{Y}^T(\mathbf{I}_N - \mathbf{Z}\mathbf{\Gamma}^{-1}\mathbf{Z}^T)\mathbf{Y} \quad (84)$$

THE MAXIMUM LIKELIHOOD ESTIMATION

Similarly to subsection , we can assume that $\epsilon \underset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Sigma$, and estimate the parameters of VAR(p) using the maximum likelihood approach. The (multivariate) negative log-likelihood function reads:

$$\begin{aligned} -\log \mathcal{L}(\beta, X) &= \frac{dN}{2}\ln(2\pi) + \frac{N}{2}\log|\Sigma| \\ &\quad + \frac{1}{2}\mathrm{Tr}\left((Y - Z\beta)\Sigma^{-1}(Y - Z\beta)^{T}\right) \end{aligned} \tag{85}$$

Hence under the Gaussian assumption, we can see that maximizing the likelihood is equivalent to minimizing $J(\beta)$ in equation (75).

THE YULE-WALKER ESTIMATION

By multiplying $x(i)$ in the first equation, with $x(i-j)$, and taking the mathematical expectation, we have :
For all $j \in \{0, \dots, p\}$:

$$\mathbb{E}\{x(i)x(i-j)^{T}\} = \sum_{k=1}^{p} \beta_k \, \mathbb{E}\{x(i-k)x(i-j)^{T}\} + \mathbb{E}\{\epsilon(i)x(i-j)^{T}\} \tag{86}$$

Recall that $S(j) = \mathbb{E}\{x(t)x(t-j)^{T}\}$ is the co-variance function defined in 9 of a stationary process x. After some manipulation, (86) becomes :

$$S(j) = \sum_{k=1}^{p} \beta_k S(j-k) \text{ if } j > 0 \tag{87}$$

$$S(0) = \sum_{k=1}^{p} \beta_k S(-k) + \Sigma \tag{88}$$

where $S(-k) = S^{T}(k)$. If we write equation (87) in large matrix format :

$$\Gamma\beta = e_1 \otimes \Sigma \tag{89}$$

where $\beta = [I_d, -\beta_1^{T}, -\beta_2^{T}, \dots, -\beta_p^{T}]$, Γ is a Block-Toeplitz matrix where each block $\Gamma_{ij} = S(i-j)^{T}$ and $e_1 \otimes \Sigma = [\Sigma, 0_d, \dots, 0_d]$.

In the univariate case ($d = 1$), the Levinson-Durbin Recursion [39] is usually used to solve this linear equation. The advantage of this recursion is that it ensures the stability of the model AR(p), and runs in

$\mathcal{O}(p^2)$. The generalization of Levinson-Durbin algorithm to the multivariate was proposed by Whittle [140] which requires the inversion of $d \times d$ matrices only. The algorithm solves this Block-Toeplitz system, called the *Modified Yule-Walker Equation* (in $\mathcal{O}(d^2p^2)$).

2.2.2 NON-LINEAR MODELS

"La caractéristique première d'une non-propriété est de ne pas être caractérisée !" [1]

The hypothesis of linearity has occupied a premier place in signal processing between 1930 and 1940 due to its inherent simplicity from conceptual and implementational points of view. However, there are many practical situations where non-linear processing of signals is needed [55]. Unlike linear systems where a unique impulse response fully characterizes the system, there exist different ways to characterize non-linearity. We have already defined one way by using higher order statistics and higher-order spectra defined in Chapter 1, and we refer the reader to [87] for more details. We define three other prominent models to tackle non-linear signals. In the sixties, started the emergence of work concerning non-linear systems, with publication of the monograph "Nonlinear problems in random theory" by Norbert Wiener. His students [127] have developed popular polynomial models known as Volterra filters. Another possible way to extend the scope of linear models to nonlinear processing is by using the kernel trick. Finally, we introduce the state-of-the-art methods based on neural networks.

KERNEL MACHINES

An elegant way to extend linear models to non-linear ones is by using the concept of *kernel machines* in ML terminology. A kernel k is a symmetric and continuous function defined on $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ where $\mathcal{X}$ is an input space.

Definition 2.2.1 (From [7]) *If* $\forall a_i, a_j \in \mathbb{R}$, $\sum_{i,j} a_i a_j h(x_i, x_j) \geqslant 0$, *the kernel corresponds to a unique inner product in an arbitrary feature space* $\mathcal{H}$. *The feature map is constructed using a kernel function* $h : \mathcal{X} \mapsto \mathcal{H}$*:*

$$k(x_i, x_j) = \langle h(x_i), h(x_j) \rangle \tag{90}$$

The main idea is to map the input space to a feature space using a non-linear function h, usually of a higher-dimension. This principle has shown its efficiency initially with Vapnik's Support Vector Machines (considered also for time-series). Using this property, authors

[78] propose straightforward approach to define a non-linear auto-regressive model:

$$h(x(n)) - \mu = \sum_{k=1}^{p} \beta_k (h(x(n-k)) - \mu) + \epsilon(n) - (1 - \sum_{k=1}^{p} \beta_k)\mu \tag{91}$$

where $\mu = \mathbb{E}\{h(x(n))\}$. Provided no guarantee that the process is still zero-mean in the feature space, the AR model is written in its mean adjusted form. The authors derive the Yule-Walker equation of the model (91) and take use of its structure by using the concept of expected (lagged) kernels to estimate efficiently the coefficients of $AR(p)$. However, one should return interpret-able predictions in the initial input space $\mathcal{X}$, the problem is this is not always guaranteed as the pre-image could be nonexistent or not unique.

pre-image problem

The kernel trick is powerful for two reasons. First, the feature mapping usually admits an implementation that is computationally efficient. Second, the parameter estimation of the non-linear model with respect to $\mathbf{x}$ is done using convex optimization techniques that are guaranteed to converge efficiently. As for the pre-image problem, it is solved using non-linear optimization techniques[78].

VOLTERRA MODELS

Another prominent class of non-linear model is based on Volterra filters. Let $x(n)$ and $y(n)$ be two signals connected by a function F such that $y(n) = F(x(n))$. The application of F to $x(n)$ involves its past and future values. Volterra model is a polynomial (with the respect to the input) with memory and anticipation.

$$y(n) = h_0 + \sum_{i=1}^{+\infty} f_i(x(n)) \tag{92}$$

$$f_i(x(n)) = \sum_{k=\infty}^{+\infty} h_i(k_1, \ldots, k_i) x(n-k_1) \ldots x(n-k_i) \tag{93}$$

whererin $h_i(k_1, \ldots, k_i)$ is termed the ith order kernel of the development. $f_i(x(n))$ is said homogeneous because it only comprises terms of the same order. It is termed homogeneous filter of order i. These definitions being presented, they involve infinite sums and thus unfeasible in practice. The p^{th} order, degree M order Volterra filter uses a finite summation $\sum_{i=1}^{p} f_i(x(n))$, $\sum_{k_i=0}^{M-1} h_i(k_1, \ldots, k_i) x(n-k_1) \ldots x(n-k_i)$.

Volterra filters have the remarkable property that they are linear with respect to $f_i(x(n))$ which simplifies the design of gradient-based and recursive least squares adaptive algorithms. Like any kernel-based method, it requires the careful selection of the kernels.

Finally, a widely used class for non-linear modeling is based on ANN. Since this class is currently state-of-the-art, we elaborate on it in more detail. For that, introducing a new terminology and relevant theory is necessary. In the following, we only introduce the building blocks necessary for understanding architectures relevant to our framework of time-series forecasting.

ARTIFICIAL NEURAL NETWORKS: RELEVANT THEORY AND TERMINOLOGY

ANN were once popular for a short time between 1940 and 1970, took a two-decade hiatus, and have been popular ever since. Their success is due to many factors, namely the Back Propagation (BP) and the computational advances. They are called *artificial* because the idea was to imitate the functioning of the biological brain. The types of networks described here are by no means the only kinds of ANN architectures found in the literature.

They comprise *units*, also called *neurons* or *nodes*. In each unit, the input undergoes a succession of multiplications by the weights of edges followed by a non-linear activation function to finally provide an output. Mathematically, this can be represented by the equation (94) where $a, f, \boldsymbol{x}, \boldsymbol{w} \overset{\text{def}}{=} [w_1, \dots, w_n]^T, b$ are respectively the output, non-linear function, the input vector, the weights vector and the neuron bias.

$$z = \boldsymbol{x}^T \boldsymbol{w} + b \tag{94}$$
$$a = f(z) \tag{95}$$

see from (94) how linear regression can be thought of as a neural network with one output and a linear activation function.

Typical activation functions are the identity function I, *rectified linear units (ReLU)*, *the sigmoid* (σ) and the hyperbolic tangent (tanh).

$$I(z) = z \tag{96}$$
$$\text{ReLU}(z) = \max(0, z) \tag{97}$$
$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{98}$$
$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{99}$$

The feed-forward neural networks comprise several units, organized in *layers*. The first layer that receives the input is termed *the input*

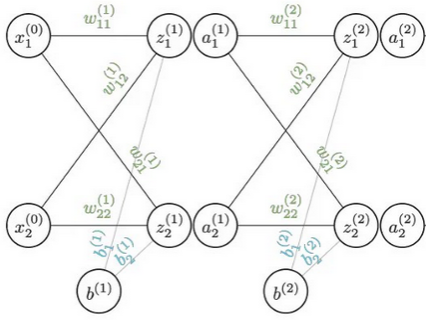
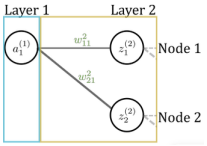


Figure 3: Illustration from [126] of a feed-forward neural network with one hidden layer. The weighted sum (z nodes) and the activations (a nodes) are split for clarity only, they are customarily considered one node.

layer, the last layer that produces the output is termed *the output layer*. The layers in between are called collectively *hidden layers*. The overall length of the chain L is termed the depth of the model. For $l \leqslant L$:

$$\begin{aligned} \mathbf{z}^{(l)} &= \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \\ \mathbf{a}^{(l)} &= f(\mathbf{z}^{(l)}) \end{aligned}$$



$\mathbf{a}^{(l)}$ *is the activation at the hidden layer* l, $w_{ij}^{(l)}$ *is the edge weight between neuron* i *and its predecessor* j.

where $\mathbf{W}^{(l)}, \mathbf{b}^{(l)}$ are respectively the weight matrix and bias vector of layer l. Finally, the output layer is obtained from the final layer:

$$\mathbf{o} = f(\mathbf{W}^{(o)}\mathbf{a}^{(L)} + \mathbf{b}^{(o)})$$

TRAINING ANN

The optimization algorithm that minimizes the error between the input and the output is usually the *Gradient descent* algorithm. This algorithm goes down to the minimum of the objective function $J(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\{\mathbf{W}^{(l)}\}, \mathbf{b}^{(l)})$, after each pass on the data. Each pass on the whole training data $(\mathbf{x}_i, y_i)_{1 \leqslant i \leqslant N}$ is termed an *epoch*. A gradient descent update at the $(t+1)$st iteration has the form:

$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} - \eta \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\beta}^{(t)}} J(\boldsymbol{\beta}^{(t)}, \mathbf{x}_i, y_i) \tag{100}$$

where η is the learning rate. If the gradient descent is performed on the whole data, then it is called *batch gradient descent*, this is the most accurate way to estimate the parameters, however the operations involved scale linearly with the number of observations making them very costly for large data. An efficient alternative is to use Stochastic Gradient Descent (SGD), where *mini-batches* of fixed size M (usually

small) are sampled from the training data to yield an unbiased estimator of the gradient of the loss w.r.t parameters.

Gradient descent involves calculating $\nabla_\beta J(\beta, x_i, y_i)$. The gradients are computed using the BP that allows the cost to flow backward through the network by using calculus chain rules. We first calculate $\frac{\partial J}{\partial W^{(o)}} = \frac{\partial J}{\partial o}\frac{\partial o}{\partial W^{(o)}}$, and then $\frac{\partial J}{\partial W^{(L)}}, \ldots, \frac{\partial J}{\partial W^{(1)}}$ working backward through the network:

$$\frac{\partial J}{\partial W^{(l)}} = \frac{\partial J}{\partial o}\frac{\partial o}{\partial a^{(L)}}\frac{\partial a^{(L)}}{\partial a^{(L-1)}} \cdots \frac{\partial a^{(l)}}{\partial W^{(l)}}$$

Some terms in the chain rule expression for the layers are shared with each other. Computationally, this means that the common terms are stored and reused rather than re-calculating the entire expression and only the terms that are particular to the current layer are calculated.

Some training issues

Note that when using tanh or sigmoid activation functions in hidden layers, these functions saturate for very small of very large values of the node z_i, i. e. σ saturates at 0 or 1, and the hyperbolic tangent at -1 or 1. Either way, their derivative is close to 0, causing the gradient to collapse to 0 (by chain rule of calculus). This is the vanishing gradient problem.

As a matter of fact, the *rectified* linear unit was proposed to alleviate this problem, since its derivative is constant at 1 for $z > 0$ and 0 otherwise. Loosely said, the gradient travels unchanged to the first layers or it becomes exactly 0 on the way. Other training issues of ANN include the influence of the initialization on the Gradient descent, the presence of local minima, overfitting due to over-parameterized networks and lack of guidelines for avoiding this problem and choosing the appropriate architecture. We shall first shed more light on this vanishing gradient problem in the context of Recurrent Neural Network (RNN). We are interested in the class of RNN because they are a subset of ANN with loops, that are suitable for dealing with sequential data i. e. time-series.

NEED FOR RECURRENT NEURAL NETWORKS FOR TIME-SERIES

Most of the neural networks are designed for i.i.d samples, and again this assumption does not hold for time-series. RNN are a type of ANNs

designed to handle sequential data. They accomplish that by maintaining a hidden state h_t that acts as a memory of past information.

$$\mathbf{h}_t = \boldsymbol{W}_{hx}\boldsymbol{x}(t) + \boldsymbol{W}_{hh}\mathbf{h}_{t-1} \tag{101}$$

$$\mathbf{o}_t = \boldsymbol{W}_{ho}\mathbf{h}_t \tag{102}$$

For simplicity, we omit the additive bias and consider the linear activation function. $\boldsymbol{W}_{hx}$, $\boldsymbol{W}_{hh}$ and $\boldsymbol{W}_{ho}$ are respectively the input-to-hidden, hidden-to-hidden and hidden-to-output weight matrices. Various design patterns exist for recurrent neural network, for example we can read the entire sequence of some length T with connections between recurrent hidden units and produce a single output. Or we can produce an output at each time-step $t \leqslant T$ as expressed in (101). Denote $j(\mathbf{o}_t, y_t)$ the cost function at time-step t, and the objective function $J(\mathbf{o}_t, y_t) = \frac{1}{T}\sum_{t=1}^{T} j(\mathbf{o}_t, y_t)$ the loss over T times-steps.

BACK PROPAGATION THROUGH TIME

The unfolded recurrent neural networks can bee seen as a multi-layer neural network, of which all layers contain the same parameters, and an algorithm termed Back Propagation Through Time (BPTT) similar to BP can be used to update the internal weights. First of all, differentiating $J(\mathbf{o}_t, y_t)$ w.r.t to the output at any time-step t reads:

$$\frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{o}_t} = \frac{\partial j(\mathbf{o}_t, y_t)}{T \partial \mathbf{o}_t} \tag{103}$$

$$\frac{\partial J(\mathbf{o}_t, y_t)}{\partial \boldsymbol{W}_{ho}} = \sum_{t=1}^{T} \frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{o}_t} \mathbf{h}_t^T \tag{104}$$

The computation of the gradients of the objective function with respect to the last hidden state h_T is straightforward. However for $t < T$:

$$\frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{h}_t} = \sum_{i=1}^{T} (\boldsymbol{W}_{hh})^{T-i} \boldsymbol{W}_{qh}^T \frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{o}_{T+t-i}} \tag{105}$$

$$\frac{\partial J(\mathbf{o}_t)}{\partial \boldsymbol{W}_{hh}} = \sum_{t=1}^{T} \frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{h}_t} \mathbf{h}_{t-1}^T \tag{106}$$

Finally, the gradient with respect to input weights reads:

$$\frac{\partial J(\mathbf{o}_t, y_t)}{\partial \boldsymbol{W}_{hx}} = \sum_{t=1}^{T} \frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{h}_t} \boldsymbol{x}(t)^T \tag{107}$$

In practice, $\frac{\partial J(\mathbf{o}_t, y_t)}{\partial \mathbf{h}_t}$ is stored to avoid duplicate calculations.

THE VANISHING OR EXPLODING GRADIENT PROBLEM

Note how equation (105) involves a potentially very large power of $\boldsymbol{W}_{hh}^{T}$ for long time-steps T. Hence, numerical instabilities occur manifesting themselves as *exploding* or *vanishing* gradient problems, depending on whether the weight is big or small. One way to address this is to truncate the time steps at a computationally convenient size. In practice, this truncation can also be effected by detaching the gradient after a given number of time steps [58]. In the following, we will see how more sophisticated sequence models such as LSTMs can alleviate this.

LONG SHORT TERM MEMORY

In the LSTM architecture, each unit in simple ANNs is replaced by a far more complex architecture called the LSTM unit or block. Mathematically, consider that we have h hidden units, the batch size is n, and the input $\mathbf{X}_t \in \mathbb{R}^{n\times d}$. An LSTM block contains three gates computed as follows:

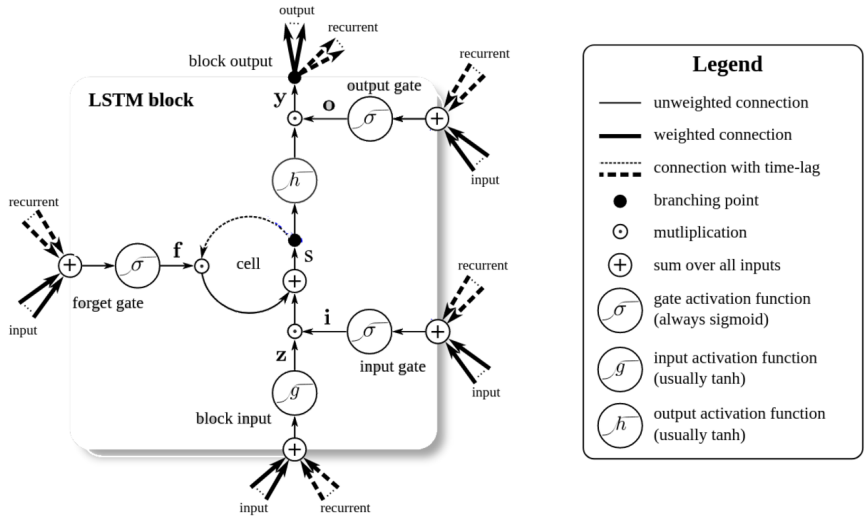


Figure 4: Image adapted from [60]

$$\mathbf{I}_t = \sigma(\mathbf{X}_t\boldsymbol{W}_{xi} + \mathbf{H}_{t-1}\boldsymbol{W}_{hi} + \mathbf{b}_i) \tag{108}$$

$$\mathbf{F}_t = \sigma(\mathbf{X}_t\boldsymbol{W}_{xf} + \mathbf{H}_{t-1}\boldsymbol{W}_{hf} + \mathbf{b}_f) \tag{109}$$

$$\mathbf{O}_t = \sigma(\mathbf{X}_t\boldsymbol{W}_{xo} + \mathbf{H}_{t-1}\boldsymbol{W}_{ho} + \mathbf{b}_o) \tag{110}$$

- Input gate $\mathbf{I}_t \in \mathbb{R}^{n\times h}$: This gate controls the connection between the input (flowing from other adjacent blocks) and the memory cell. It outputs a value of 0 or 1.

- Forget gate $\mathbf{F}_t \in \mathbb{R}^{n\times h}$: Note that when the forget takes as value 1, then the memory cell **remembers** and does not discount past information.
- Output gate $\mathbf{O}_t \in \mathbb{R}^{n\times h}$: The output gate has the same form of the forget and the input gate.

The input gate governs how much we take new data into account via the input node $\mathbf{Z}_t \in \mathbb{R}^{n\times h}$ and the forget gate $\mathbf{F}_t$ addresses how much of the old cell $\mathbf{S}_{t-1}$ we retain.

Last, we need to define the output of the memory cell, i. e. the hidden state $\mathbf{H}_t$:

$$\mathbf{Z}_t = \tanh(\mathbf{X}_t W_{xc} + \mathbf{H}_{t-1} W_{hc} + \mathbf{b}_c) \tag{111}$$
$$\mathbf{S}_t = \mathbf{F}_t \odot \mathbf{S}_{t-1} + \mathbf{I}_t \odot \mathbf{Z}_t \tag{112}$$
$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{S}_t) \tag{113}$$

Loosely said, the input and forget gates give the model the flexibility to decide when to change the value of the cell $\mathbf{S}_t$ as a response to subsequent inputs. When back-propapagating the gradient at various times-steps $t \geqslant 1$ (BPTT), the derivative of the cell state $\frac{\partial s_t}{\partial s_{t-1}}$ solves the vanishing gradient problem. Its additive properties and changing values make the gradient less likely to be degenerate, unlike RNN where the same weights are back-propagated through long time-steps.

OTHER ARCHITECTURES

DEEP LSTMS

Deep learning consists of stacking multiple neural networks. The premise is that each layer will receive the previous output, and then performs a learning on it, to construct more and more complex features. This is a new pipeline to construct features, completely different from traditional machine learning where the features were handcrafted. Deep learning methods are now state-of-the-art in many fields from computer vision, image processing to language translation. This success is owed to availability of large labeled datasets and advances made in computer engineering However, despite the intuition that deeper architectures would yield better results than "shallow" ones, empirical tests with deep networks had found similar or even worse results when compared to networks with only one or two layers [137].

RNNs can be thought of as Deep neural network that comprise T hidden layers. The aim of this *depth* is not however to extract abstract features, but simply to keep in memory the long-range dependencies.

Therefore, one can think of cascading LSTMs, the first layer will handle the dynamic evolution of the raw data, and the next layer will receive as output $\mathbf{H}_t$ of each sequences, and re-perform the learning on the hidden states, and so on for each stacked LSTM to provide hierarchical abstract features of the evolution of data. A two-layer architecture of stacked LSTMs has been used in [95] for anomaly detection in time-series.

ECHO-STATE NETWORKS

Echo-State Network (ESN) were originally proposed by [74] for time-series forecasting in wireless communication. They were designed to mitigate the numerical instabilities of RNNs by eliminating the need to compute the gradient with respect to the hidden layers. The core of ESN is a sparsely connected random RNN called a *reservoir*. The weights of the reservoir are not learned via gradient descent but rather fixed. The reservoir must satisfy the *Echo state property*[143]. Only the hidden to output weights are tuned via least squares for prediction tasks.

HYBRID ARCHITECTURES

Recently, "traditional" time-series analysis methods and machine learning are starting to merge through the use of hybrid architectures. For example [82] presented a model for time-series forecasting using Auto-Regressive Integrated Moving Average (ARIMA) and ANN. Such models are generally constructed in a sequential manner, with the ARIMA model first applied to the original time series to capture the linear component, and then its residuals (containing non-linear relationships) are modeled using neural networks. An example of a real-world application of this hybrid modeling can be found in [49]. This holds the promise of a best-of-wo-worlds scenario where the statistical learning methods and the newly data-driven architectures are merged in a favourable way.

CONVOLUTIONAL NETWORKS FOR TIME-SERIES

Motivated by their success for classification tasks for images, researchers have started adopting Convolutional Neural Network (CNN) for time series analysis. A convolution can be seen as applying and sliding a

filter over the time series of size N. For a convolution centered around t, we have:

$$c_t = f(w \odot x(t - l/2 : t + l/2) + b) \quad \forall t \in [1, N] \tag{114}$$

Where c_t denotes the output of a convolution applied on a univariate[1] time-series x of size N with a filter w of length l. Cascading the same filter on c_t will result in a multivariate time-series whose dimension is equal to the number of filters used. The promise of using multiple filters is that of deep learning, to learn abstract discriminative features. The convolutions are usually followed by *average* or *max* pooling layers to reduce the size N of the time-series by aggregating over a sliding window. This technique allows for achieving some translation invariance of the learned features. Finally, when the layers become small enough, it is common to have fully connected layers before the output layer.

The scattering transform discussed in Chapter 1,1.3.4 can bee seen as a CNN with two main distinctions:

- The filters are fixed from a family of wavelets and not learned as in traditional CNNs
- Each hidden layer provides an output

In light of the success of CNN on image classification, they are typically used on two-dimensional data for image recognition by means of 2D filters. In [27], authors have extended the scattering transform to images using a suitable family of directional wavelet family. As a matter of fact, we have been directly dealing with available observations in the time-domain. Another paradigm would be to "transform" time-series to images; two of the most widely used transformations or mappings are recurrence plots [64], and time-frequency transformation by means of the spectrogram for example, in 1.3.3. For example, authors in [129] leverage the theoretical guarantees of the scattering transform introduced in Chapter 1,1.3.4, and the flexibility of CNN to perform unsupervised classification on time-series. More deep learning methods designed for time-series classification can be found in [73].

A WORD ON DETECTION

In this chapter, we have mainly focused on the problem of time-series forecasting because it relates to the application we have in mind, namely

1 For a d-variate time-series the filter will also be d-dimensional.

the detection problem (or anomaly detection) that will be the core of Chapter 5 and Chapter 6. The detection problem could be considered the same as finding regions in the time-series for which the forecasted values are too different from the actual ones. How to quantify *too different* while providing theoretical guarantees on the detection rates is the core subject of Chapter 4.

In the final section, we discuss important questions that naturally arise in any learning framework. For example, the ever-increasing amount of challenging real-data promotes the use of equally complex models, when in fact it *should* not; growing the complexity of the model will not necessarily lead to the adequate model, this search of the optimal model is ruled by the principle of *parsimony* discussed in the following.

2.3 PARSIMONY: ADEQUACY VS. COMPLEXITY

The generalization error vs. fit error is reminiscent of the bias-variance trade-off in the estimation framework

In deciding which model is the best, criteria that allow model comparisons are necessary. A major tenet to follow is *parsimony*, mostly known now as *Occam's razor*, that is the model with the smallest possible number of parameters is preferred. First, let's consider the VAR(p) where p has been supposed known. In practice, this is not the case. If we oversimplify the model by choosing $\hat{p} < p$, the approximations made by least squares will perform poorly, even on the training data. An unnecessarily complex model with $\hat{p} > p$ parameters will perform well (overfit) on training data but will have poor performance on unseen data (generalization error). This search for $\hat{p}$ can be formulated mathematically by minimizing an objective function of the model's complexity subject to the constraint of model adequacy. Parsimony is usually described as a function of degrees of freedom such that fewer parameters (thus higher degrees of freedom) correspond to more parsimony.

The Minimum Description Length (MDL) proposed by [120] is connected to Occam's rasor in ML. It quantifies the complexity of the model by the length of the code obtained when that model is used to compress the data. Loosely said, if the model was successful at learning the data, then it can compress it using a short code. For further details on this theory, we refer the reader to Chapter 14 of [34].

If the fitting is carried out by maximization of log-likelihood $\log \mathcal{L}$ (ensuring that maximum likelihood estimators are asymptotically Gau-

sian). The most prominent Akaike, Bayesian and Hannan-Quinn Information Criteria (IC) can be used:

$$\hat{p} = \underset{p}{\operatorname{argmin}}\, \mathrm{IC}(p) \tag{115}$$

$$\mathrm{AIC}(p) \overset{\text{def}}{=} -2\log\mathcal{L} + 2(\#\text{ parameters}) \tag{116}$$

$$\mathrm{BIC}(p) \overset{\text{def}}{=} -2\log\mathcal{L} + \log(N)(\#\text{ parameters}) \tag{117}$$

$$\mathrm{HQIC}(p) \overset{\text{def}}{=} -2\log\mathcal{L} + 2\log\log(N)(\#\text{ parameters}) \tag{118}$$

For a Gaussian d-dimensional VAR(p) process, we can write them as [94]:

$$\mathrm{AIC}(p) = \log(|\hat{\boldsymbol{\Sigma}}|) + 2\frac{d^2p}{N} \tag{119}$$

$$\mathrm{BIC}(p) = \log(|\hat{\boldsymbol{\Sigma}}|) + \frac{\log(N)d^2p}{N} \tag{120}$$

$$\mathrm{HQIC}(p) = \log(|\hat{\boldsymbol{\Sigma}}|) + 2\frac{\log\log(N)d^2p}{N} \tag{121}$$

Unlike $\mathrm{AIC}(p)$, $\mathrm{BIC}(p)$ and $\mathrm{HQIC}(p)$ have the advantage of being consistent estimators, meaning that the probability of selecting the true lag length approaches 1 as the sample size goes to infinity. The MDL approach gives a selection criterion formally identical to the BIC, however recall that it is motivated from an optimal coding viewpoint.

In general, there exists a myriad of procedures for choosing a model via other selection criteria or statistical hypothesis testing such as the Likelihood Ratio test [115].

There also exists a class of these procedures that does not rely on any probabilistic assumptions. It is based instead on data resampling, the most popular is *cross-validation* (with its many variants such as k-fold or leave out one cross-validation)[135]. These procedures are based on the idea of repeating the training and testing computation on different randomly chosen splits of the original training dataset to have an estimate of the generalization error of the learner.

NECESSITY TO RETHINK GENERALIZATION FOR NEURAL NETWORKS

Which network size is more appropriate for a given problem? Unlike linear models, the answer to this question is not straightforward.

There is a number of theoretical results concerning the number of hidden layers in a network, for example Hornik [72] has shown that a single hidden layer feed-forward network with as few as one hidden layer and arbitrary (bounded and non-constant) activation functions

can approximate any function of interest to any desired degree of accuracy. Provided enough hidden units are available of course. But how to choose the number of hidden units? Despite neural networks being complex in nature, following the principle of parsimony has both theoretical and practical advantages: Smaller networks require less hardware implementation costs, training a smaller network require less expensive computations and produce fast propagation delays from the input to the prediction. Most importantly, they have good expected generalization capabilities.

The usual approaches pursued to impose parsimony are heuristics such as *regularization*, *stopped training* and *pruning methods*. The underlying idea of these approaches is that we allow the learner to be over-parameterized but we force the parameters to be sparse either by adding a penalty term to the cost function (for example, L^1 regularization to enforce sparsity or L^2 to favour low magnitude solutions, or both as in ElasticNet [146]); or stopping the learning when the error on the validation set starts to grow, or removing the less significant weights by pruning. A survey of the latter is provided in [119]. The main disadvantage of regularization, pruning and stopped training is that these methods comprise of a strong judgemental component[5] (choice of the penalty term, the stopping criterion, the significance criterion in pruning), which makes the model building process difficult to reconstruct. Therefore, attempts have been made to apply hypothesis statistical testing to ANNs [5], or derive Information criteria for a feedforward ANNs [109]. Research on model selection in neural networks is still an open problem; the Probably Approximately Correct (PAC) learning has drawn together statistics and ML in an attempt of deriving mathematical foundation for the applied practice of ML[65].

Beyond enhancing generalization capabilities, imposing parsimony in neural networks is of crucial importance as larger networks translate to greater computer demands, and by extension to greedy energy costs. Over-parameterized ANNs are also culprit of their lack of explain-ability. Rethinking how we do deep learning in the context of climate change and making it user-friendly are a necessity. Some paradigms promise to evade computational burden by *transfer learning* and *fine-tuning* where trained labels are re-used rather than starting the training from scratch or *knowledge distillation* techniques [59].

2.4 CONCLUSION

We started the introduction of Part i with an important question: *"For what class of problems is a method intended to be used?"*. The confusion about this question is a result of failing to define the problem explicitly

enough. Any available a priori knowledge must be exploited and may lead to the preference of a particular model. If suitable to the problem at hand (given that all underlying hypotheses are met by the data), choosing a linear model comes with its advantages: Guarantees of convergence to optimal weights, consistent estimates of the number of parameters and various statistical procedures to test the model's adequacy. If judged too simplistic and biased for the problem, resorting to non-linear models allows for learning more complex mappings, provided that large labeled datasets are available but at the expense of losing convergence guarantees, expensive computational complexity and lack of explainability.

	Time (T)/ Frequency (F)	Linear (L)/ Non-linear (NL)	Stationary (S)/ Non-stationary (NS)
(V)ARMA	T	L	S
(V)ARIMA	T	L	NS
Fourier Transform	F	L	S
Spectrogram	F	L	NS
Scattering Transform	F	L	NS
Kernal AR	T	NL	S
Volterra models	T	NL	S
RNNs-LSTMs	T	NL	S

Table 1: Synoptic of the presented tools in Part i. Parametric and non-parametric models presented are grouped, the columns intend to classify the different tools in time or frequency procedures, for linear or non-linear systems, procedures for stationary or non-stationary processes.

This discussion closes Part i which was dedicated to the introduction of the main technical tools related to the main contributions of this manuscript. Hopefully, this part has shown how Gaussianity occupies a premier place in signal processing and statistical learning. Hence, validating this assumption is necessary. In the following chapter, we proceed with the first contribution focused on a statistical procedure to assess the departure from Gaussianity. Subsequently, in Chapter 5, Chapter 6 this contribution will be translated to an operational detector tested on both synthetic and real data.

Part II

JOINT NORMALITY TEST FOR TIME-SERIES

This part is devoted to the presentation of our contributions to testing Gaussianity of multivariate time-series. First, Chapter 4 provides a bird's eye view on existing procedures and explains the necessity of deriving a new test in our framework. Chapter 4 details the theoretical background and calculation steps necessary for deriving the test, our main contribution will be stated as a theorem which then will be implemented and tested on colored copula. Successively, Chapter 5 efficiently translates our theoretical results into a practical sequential algorithm for the detection of non-Gaussian signals embedded in Gaussian noise.

3

BIRD'S EYE VIEW ON NORMALITY TESTS

> *«Amends might be made in the interest of the new generation of students by printing in leaded type in future editions of existing text-books and in all new text-books:*
> *Normality is a myth; there never was, and never will be, a normal distribution.»*
>
> Geary, 1947

Geary's over-statement is meant to point out situations where Gaussianity is assumed without any foundation. When assigned to a population, it gives birth to many desirable properties, but as important as the assumption of Gaussianity is, it is equally important to verify its actual validity. There is a large literature on normality tests that we cannot cover due to the continuous proliferation of the procedures. However we propose instead a taxonomy based on two criteria, along with some selected examples. The first aim is to highlight the imbalance between classes of normality tests, and motivate the necessity of deriving a novel normality test in our framework.

Contents

3.1 PROBLEM STATEMENT

Normality tests are a subset of binary hypothesis testing procedures:

> **Problem P1:** Given a finite sample of size N of d-variate random variables $\mathbf{x}(n)$, $\mathbf{X} \overset{\text{def}}{=} \{\mathbf{x}(1), \ldots, \mathbf{x}(N)\}$:
>
> $$\mathcal{H}_0 : \mathbf{X} \text{ is Gaussian} \quad \textit{versus} \quad \bar{\mathcal{H}}_0 \tag{122}$$

where variables $\mathbf{x}(n) \in \mathbb{R}^d$ are identically distributed. We do not oppose any alternative to the null hypothesis $\mathcal{H}_0$. This is important since it implies two things, first that the test statistic needs to be fully described under the null hypothesis only. Second, there cannot be an optimal test for the null hypothesis, the type II error remains undefined and only one parameter controls the level of the test, that is the significance level α, false alarm rate or the type I error in statistics:

$$\alpha = P(\text{choose } \bar{\mathcal{H}}_0 | \mathcal{H}_0 \text{ is true}) \tag{123}$$

If the test does not lead to the rejection of the null hypothesis, the latter is by no means confirmed or validated. Myriad of test procedures have been derived and in the following we propose a taxonomy based on whether they were designed for scalar processes ($d = 1$) or $d > 1$, and whether they assume that $\mathbf{x}(n)$ is a white process ,i. e. $\mathbb{E}\{\mathbf{x}(n)\mathbf{x}(m)^T\} = 0$ for $n \neq m$ or a colored (as opposed to white) process ,i. e. identically distributed but non-independent.

Before starting the taxonomy of normality tests, let's define some desirable properties a normality test should satisfy:

Some desirable properties of the test statistic

- Affine invariance: Any test statistic $t(\mathbf{X})$ should satisfy i. e.

  $$t(\sum_n \mathbf{A}_n \mathbf{x}(n) + \mathbf{c}) = t(\sum_n \mathbf{x}(n))$$

  for any d by d matrix $\mathbf{A}_n$ and $\mathbf{c} \in \mathbb{R}^d$. Otherwise, the test could have conflicting conclusions on the same data.

- Fully defined limiting distribution under the null hypothesis

- In his critical review, Henze [67] points out the fact that skewness is inconsistent against non-normal elliptically symmetric distributions and Kurtosis-based tests are only consistent if the population kurtosis of the alternative is different than that of

normal distribution. Consistency is unarguably a desirable property of a test statistic, we will be less strict in demanding this property from the tests we choose to elaborate upon. Most of the normality tests fail to be consistent against *all* alternatives. Any a priori knowledge about the data should be used to choose an adequate descriptive measure.

- Feasibility with respect to the number of samples and their dimension. We will only deal with the case $d < N$. The reader interested in high-dimensional settings $d > N$ is referred to [90] where authors propose a projection-based test for high dimensional setting when N is small and in the recent work of [142] a test statistic based on Mardia's Skewness is defined in the asymptotic framework $N, d \to \infty$.

3.2 A TAXONOMY OF NORMALITY TESTS

SOME GRAPHICAL TOOLS

A good preliminary approach to assess the normality of observations is to use graphical tools.

THE HISTOGRAM

For simplicity consider $x_1, \dots, x_N \in [0, 1]$ i.i.d observations. The histogram splits the set $[0, 1]$ into bins and uses the count of values in the bin as a density estimate. If we have k bins, this yields the partition:

$$B_1 = [0, \frac{1}{k}), B_2 = [\frac{1}{k}, \frac{2}{k}) \dots, B_k = [\frac{k-1}{k}, 1]$$

Many rules dictate the choice of the number of bins, such as the square root choice $k = \lfloor \sqrt{N} \rfloor$ or Rice rule $k = \lfloor 2N^{3/2} \rfloor$ to name a few. Following the arguments in [53], it is desirable that the number of bins is proportional to the cube root of the sample size. For $x \in B_k$:

$\lfloor . \rfloor$ *denotes integer part*

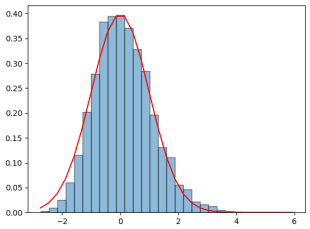
$$\hat{f}(x) = \frac{k}{N} \sum_{i=1}^{N} \#(x_i \in B_k)$$

The histogram is a popular density estimator as it is easy to draw. It is usually smoothed using a Kernel density estimator defined as

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} k(\frac{x_i - x}{h})$$

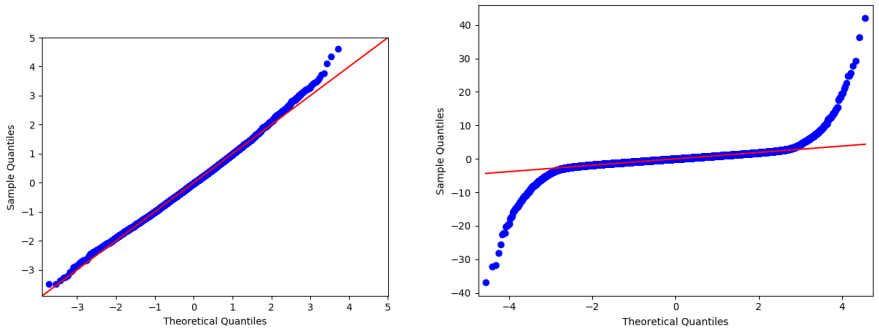
where h controls the "bandwidth" of the smoothing.

In blue histogram of a standard normal variable. In red, the kernel density estimator.



QUANTILE-QUANTILE PLOTS

The Q-Q plot of sample x is a plot of the order statistics $x_{(1)} \leqslant x_{(2)} \leqslant \ldots x_{(N)}$ from a distribution F against $y_{(1)} \leqslant y_{(2)} \leqslant \ldots y_{(N)}$ of a specified matching distribution G. If F matches G then for some $t \in \mathbb{R}$ and $s > 0$, $F(x) = G(\frac{x-t}{s})$, translating to $(\mathbb{E}\{x_{(i)}\} - t)/s = \mathbb{E}\{y_{(i)}\}$ which means that the order statistics will be centered on the line $y = \frac{x-t}{s}$.



(a) Normal distribution (b) Lepokurtic distribution

Figure 5: QQ-plot of observations against $\mathcal{N}(0, 1)$

Detailing two graphical approaches is by no means exhaustive, we refer the reader to the survey by Fisher [50]. The extension of univariate graphical tools to multivariate settings has been suggested by many researchers e. g.[41],[102].

For the sake of simplifying some of the presented test statistics, consider $y_i = \frac{x_i - \mu}{\sigma}$ in the scalar case, and $\mathbf{y}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \mu)$ for $\mathbf{x}_i \in \mathbb{R}^d$.

3.2.1 FOR SCALAR I.I.D PROCESSES

A class of normality tests, of which we cite 2, relies on measuring the distance between the empirical cumulative distribution of the i.i.d observations and F the cumulative distribution function of a standard normal.

KOLMOGOROV-SMIRNOV TEST

$$\begin{aligned} K &= \sup|\hat{F}(y) - F(y)| \\ K &= \max(D^+, D^-) \\ D^+ &= \max_{i\in\{1,\dots,N\}} \left(\frac{i}{N} - F(y_i)\right) \\ D^- &= \max_{i\in\{1,\dots,N\}} \left(F(y_i) - \frac{i-1}{N}\right) \end{aligned}$$

Kolmogorov derived the asymptotic distribution of K, and Smirnov gave percentage points of this statistic in 1948, and only then started the practical use of this statistic. A practical problem is that usually the mean and variance are unknown and estimated from the observations, in this case the asymptotic distributions of K and $\hat{K}$ differ substantially.

Another test that's based on the distance between the empirical function is Anderson-Darling [6] test statistic:

$$AD = -N - \sum_{i=1}^{N} \frac{(2i-1)}{N} \left(\log F(y_i) - \log F(y_{N+i-1})\right) \quad (124)$$

Using a bootstrap technique, this test has been extended to colored processes by [118] as we will see in the subsection 3.2.3.

THE SHAPIRO-WILK TEST

The test statistic is the ratio between the linear estimation of σ^2 based on order statistics and the empirical variance [131]:

$$W = \frac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)^2}{\sum_i (x_i - \mu)^2} \quad (125)$$

such that $x_{(1)} \leqslant x_{(2)} \leqslant \dots \leqslant x_{(N)}$ are order statistics and the coefficients $\mathbf{a}^T = \frac{\mathbf{m}^T \mathbf{V}^{-1}}{\mathbf{m}^T \mathbf{V}^{-2} \mathbf{m}}$, $\mathbf{m} = [m_1, m_2, \dots, m_N]$ are the means of order statistics of a Gaussian i.i.d random variables.[1]

This test is suitable for small sample size ($N \leqslant 100$). Another test of the same kind as Shapiro is proposed by D'Agostino et al (1971) [36]. Shapiro and Francia [131] propose a statistic for large samples.

SKEWNESS-KURTOSIS TEST

This test is based on two descriptive measures: Skewness and Kurtosis. The omnibus test initially proposed by D'Agostino-Pearson [37]

1 The vector of order statistics for $\mathcal{N}(0, \sigma^2)$ is $\mathbb{E}\{x_{(1)} \dots x_{(N)}\} = \mu + \mathbf{m}\sigma$

combines the standardized skewness and kurtosis, however this test assumes independence between the two moments, which is not the case and thus is not recommended in practice. Bowman and Shenton [21] proposed an improvement of the test statistic by comparing:

$$SK = \frac{N}{6}\left(\frac{\hat{\mu}_3}{\hat{\sigma}^{3/2}}\right)^2 + \frac{N}{24}\left(\frac{\hat{\mu}_4}{\hat{\sigma}^2} - 3\right)^2 \tag{126}$$

to the upper critical values of the asymptotic *chi*-distribution with 2 degrees of freedom χ^2_2. This statistic is justified by this limiting result for independent data under the null hypothesis:

$$\sqrt{N}\begin{pmatrix} \frac{\hat{\mu}_3}{\hat{\sigma}^{3/2}} \\ \frac{\hat{\mu}_4}{\hat{\sigma}^2} \end{pmatrix} \to_d \mathcal{N}\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right) \tag{127}$$

This test statistic was subsequently derived by Jarque and Bera [75] as the Lagrangian Multiplier (LM) test against the Pearson family distributions. It's also called the Jarque-bera test[2]. Additionally, [21] proposed another statistic:

$$SK' = X_s^2\left(\frac{\hat{\mu}_3}{\hat{\sigma}^{3/2}}\right) + X_s^2\left(\frac{\hat{\mu}_4}{\hat{\sigma}^2}\right)$$

where $X_s^2()$ is the normal variable obtained by Johnson's S_u transformation. Since $X_s\left(\frac{\hat{\mu}_3}{\hat{\sigma}^{3/2}}\right)$ and $X_s\left(\frac{\hat{\mu}_4}{\hat{\sigma}^2}\right)$ are normal and nearly independent, this statistic has also a Chi-squared limiting distribution.

The rationale behind using an *omnibus* test is to overcome the shortcomings of inconsistency when deriving a test based on only one measure. The list of normality tests for scalar i.i.d process can go on and on. For completeness, see excellent survey of [99] where at least 50 univariate tests were listed.

3.2.2 FOR d-DIMENSIONAL WHITE PROCESSES

Many generalizations of the univariate test procedures presented in the previous section for testing multinormality can be found in the literature. See the survey of Henze [67], and more recently by Henze & Ebner [42]. The latter focuses on affine invariant and consistent test procedures. A complementary survey relaxing theses properties can be found in [102].

2 Its implementation in Python statistical package goes under the name of Jarque-bera.

MARDIA'S SKEWNESS-KURTOSIS TEST

Recall the multivariate generalizations of the Skewness $\beta_{1,d}$and Kurtosis $\beta_{2,d}$ defined in Equation 28, defined in Chapter 1, and their scalar counterparts Equation 31:

$$\hat{B}_{1,d} = \frac{1}{N}\sum_{i,j=1}^{N}\left(\mathbf{x}(i)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(j)\right)^3$$

and

$$\hat{B}_{2,d} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}(n)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(n)\right)^2$$

Theorem 2 (Asymptotic distribution of $\hat{B}_{1,d}$) *Under the assumption that* $\mathbf{x} \underset{\text{i.i.d}}{\sim} \mathcal{N}_d(0, \mathbf{S})$, $\frac{N}{6}\hat{B}_{1,d}$ *has a Chi-squared* χ^2 *distribution with* $d(d+1)(d+2)/6$ *degrees of freedom. Additionally,* $\mathbb{E}\{\hat{B}_{1,d}\} = \frac{d(d+1)(d+2)}{2}$ *and* $\mathbb{V}\mathrm{ar}\{\hat{B}_{1,d}\} = \frac{12(d+1)(d+2)}{N^2}$.

Note that in testing univariate normality, more emphasis has been placed on the use of $\sqrt{\hat{B}_{1,d}}$ rather than $\hat{B}_{1,d}$. Since $\sqrt{\hat{B}_{1,d}}$ is always positive, it cannot be seen as a generalization of $\mathcal{K}_1$.

Theorem 3 (Asymptotic distribution of $\hat{B}_{2,d}$) *Under the assumption that* $\mathbf{x} \underset{\text{i.i.d}}{\sim} \mathcal{N}_d(0, \mathbf{S})$, $\hat{B}_{2,d}(N)$ *is asymptotically normal, with mean* $d(d+2)\frac{N-1}{N+1}$ *and variance* $\frac{8d(d+2)}{N} + o(\frac{1}{N})$.

Hence to test multinormality, one could separately test that population skewness $\beta_{1,d} = 0$ and population kurtosis $\beta_{2,d} = d(d+2)$ by means of the theorems 2,3. Their full proofs are derived in Sections 2 and 3 of [98]. An alternative demonstration of the limiting distribution of $\hat{B}_{2,d}$ can be found in [86].

In [100], Mardia & Foster derived 6 omnibus normality tests combining the skewness and kurtosis. In their attempt to derive a similar omnibus test to the univariate case. Mardia and Foster [100] proposed three transformations of $\hat{B}_{1,d}$, and in total 6 different omnibus tests of which three account for the non-negligible correlation between $\hat{B}_{1,d}$ and $\hat{B}_{2,d}$ that is not negligible even for large N.

3.2.3 FOR SCALAR COLORED PROCESSES

There are no guarantees ensuring that the above tests are still consistent when the assumption of independence between samples (whiteness) is violated. Gasser [54] and Moore [103] both studied the effect

of dependence of time-series on the Chi-squared test and concluded that dependence yields a loss of apparent normality.

EPPS TEST

Epps [48] proposed a test for stationary time-series that tests whether the characteristic function coincides with the Gaussian's characteristic function at certain points on the real line. More formally, let $N > 0$ and

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N, \quad 0 < \lambda_i \leqslant \lambda_{i+1}\}$$

. As pointed out by many authors, the implementation of the procedure is rather difficult [92, 111, 118]. But in principle, given that the process verifies a set of conditions [48], the procedure consists of defining a sensing functions:

$$\hat{g}(x(t), \lambda) = [\cos(\lambda_1 x(t)), \sin(\lambda_2 x(t)) \ldots \sin(\lambda_N x(t))]^T$$

Additionally define:

$$\hat{g}(\lambda) = \frac{1}{n} \sum_{t=1}^{n} [\cos(\lambda_1 x(t)), \sin(\lambda_2 x(t)) \ldots \sin(\lambda_N x(t))]^T$$

$$g_\theta(\lambda) = [Re(\Phi_\theta)(\lambda_1), Im(\Phi_\theta)(\lambda_1) \ldots]^T$$

where $Re(\Phi_\theta)(\lambda_1)$ and $Im(\Phi_\theta)(\lambda_1)$ denote respectively the real and imaginary parts of the characteristic function of a normal variable Φ_θ with $\theta = (\mu, \sigma^2)$.

The joint spectral function of the sensing functions $\{g(x(t), \lambda)\}$ is estimated at frequency $\omega = 0$. Then its inverse $\mathbf{G}^{-1}$ is calculated. The test statstic proposed by Epps is then defined as the minimum of the quadratic form:

$$Q_n(\lambda) = \underset{\theta}{\operatorname{argmin}} \, (\hat{g}(\lambda) - g_\theta(\lambda))^T \, \mathbf{G}^{-1} \, (\hat{g}(\lambda) - g_\theta(\lambda))$$

Theorem 4 (Epps Test) *If $\mathbf{x}$ is a stationary Gaussian process then $nQ_n(\lambda)$ has a χ^2 limiting distribution with $(2N - 2)$ degrees of freedom $\forall \lambda_i \in \Lambda$.*

As pointed out by [111], The Epps test is not consistent if the real line coincides with that of a Gaussian distribution. They alleviate this by selecting the set of λ randomly. Then they incorporate this upgraded Epps test in their random projections based Gaussianity test [111] that will be discussed in subsection 3.2.3.

CORRECTION OF SKEWNESS-KURTOSIS TEST BY GASSER

Let $S(j) = \mathbb{E}\{x(n)x(n+j)\}$ and $F^{(k)} = \sum_{j=-\infty}^{+\infty} S(j)^k$. The stationary process should satisfy the weak-dependent condition

$$\sum_{j=-\infty}^{+\infty} S(j) < \infty$$

Gasser [54] proposed the following limiting distribution of the Skewness and kurtosis for colored processes:

$$\sqrt{N}\begin{pmatrix} \frac{\hat{\mu}_3}{\hat{S}^{3/2}} \\ \frac{\hat{\mu}_4}{\hat{S}^2} \end{pmatrix} \rightarrow_d \mathcal{N}(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6F^{(3)} & 0 \\ 0 & 24F^{(4)} \end{pmatrix}) \tag{128}$$

However, he did not provide any formal analysis or any recommendation about the selection of the truncation number of the infinite sums.

CORRECTION OF SKEWNESS-KURTOSIS TEST BY LOBATO-VELASCO

In the same spirit, Lobato & Velasco [92] propose the following test statistic: Let

$$\hat{F}^{(k)} = \sum_{\tau=1}^{N-1} 2\hat{S}(\tau)(\hat{S}(\tau) + \hat{S}(N-\tau))^{k-1} + S^k \tag{129}$$

$$\hat{G}_k = N\frac{(\hat{\mu}_3)^2}{6F^{(3)}} + N\frac{(\hat{\mu}_4 - 3\hat{S}^2)^2}{24\hat{F}^{(4)}} \tag{130}$$

Theorem 5 (Lobato and Velasco SK test, 2004) *Let* $\mathbf{x} = [x(1), \ldots, x(N)]^T$ *be an ergodic stationary process. If* $\mathbf{x}$ *is Gaussian, then*

$$\hat{G}_k \underset{N\rightarrow\infty}{\rightarrow} \chi^2_2$$

$\hat{G}_k$ *diverges to* ∞ *whenever* $\mu_3 \neq 0$ *and* $\mu_4 \neq 3S^2$ *if* $\mathbb{E}\{x(n)^{16}\} < \infty$ *and a set of conditions defined in [92].*

RANDOM PROJECTIONS

First we need to introduce the following definition and theorem: Let $\mathcal{H}$ denote a separable Hilbert space.

Definition 3.2.1 (Dissipative distribution) *A dissipative distribution* η *generalizes an absolutely continuous distribution to the infinite dimensional space.*

Theorem 6 (Cuesta-Albertos et al.[111]) *Let* η *be a dissipative distribution on* $\mathcal{H}$ *and* $\mathbf{h}$ *and* $\mathcal{H}$*-valued random element, then* $\mathbf{x}$ *is Gaussian if and only if:*

$$\eta(\mathbf{h} \in \mathcal{H} : \langle \mathbf{x}, \mathbf{h} \rangle \text{ has a Gaussian distribution}) > 0 \tag{131}$$

Since η *is a dissipative distribution, the* $0-1$ *law holds [111] and* $\mathbf{x}$ *is not Gaussian if and only if:*

$$\eta(\mathbf{h} \in \mathcal{H} : \langle \mathbf{x}, \mathbf{h} \rangle \text{has a Gaussian distribution}) = 0 \tag{132}$$

and $\mathbf{x}$ *is Gaussian if and only if:*

$$\eta(\mathbf{h} \in \mathcal{H} : \langle \mathbf{x}, \mathbf{h} \rangle \text{has a Gaussian distribution}) = 1 \tag{133}$$

Loosely said, to test the Gaussianity of $\mathbf{x}$, we have to sample at random $\mathbf{h} \in \mathcal{H}$ using a dissipative distribution, and test whether the projection $\langle \mathbf{x}, \mathbf{h} \rangle$ is Gaussian. If the latter is Gaussian, then Gaussianity of $\mathbf{x}$ is ensured with probability 1. In practice, $\mathbf{h}$ is drawn with a stick-breaking process that makes use of beta distributions [111].

Once $\mathbf{h}$ has been fixed, the new process $\mathbf{y}^h$ can be constructed:

$$y^h(t) = \sum_{i=1}^{\infty} h(i)x(t-i) \tag{134}$$

In practice, testing for the normality of $\mathbf{y}^h$ is of course left to the practitioner, but the authors have used the Lobato-Velasco and Epps test. Even though, Theorem 6 implies that it one projection suffices to conclude, the authors advise to take more than one-projection, applying the test and then mixing the p-values using the False Discovery rate as proposed in Benjamini-Yuketieli.

BOOTSTRAP APPROXIMATION OF THE ANDERSON-DARLING TEST

Arguing that making use of classical asymptotic inference for the Anderson darling statistic is problematic and involved for time-series, authors of [118] use an auto-regressive sieve bootstrap to estimate its distribution for time-series. $x(t)$ is a stable invertible auto-regressive process expressed by:

$$x(t) = \sum_{i=0}^{\infty} \beta_i x(t-i) + \epsilon(t)$$

The main idea is to generate bootstrap sample $\epsilon(t)^B$ to approximate the residuals with a finite order auto-regressive model AR(p) by least squares estimation. From which a bootstrap sample of $x(t)^B$ is deduced using $x^B(t) = \sum_{i=0}^{p} \hat{\beta}_i x(t-i) + \hat{\sigma}\epsilon(t)^B$. They are plugged in the test statistic, and this second step is repeated M times yielding $AD_1^B, \ldots, AD_M^B$. They serve as an estimate of the distribution of AD under the null hypothesis.

HINICH'S BISPECTRUM TEST

The r^{th} order Higher-order spectrum vector defined in Equation 45 reduces in the scalar case to:

$$p_2 = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} \kappa_3(\tau_1, \tau_2) \exp^{-j\omega_1\tau_1 - j\omega_2\tau_2}$$

Given the symmetries of P_2 its principle domain is the triangular set $W = \{0 \leqslant \omega_1 \leqslant \pi, 0 \leqslant \omega_2 \leqslant min(\omega_1, 2(\pi - \omega_1))\}$.

If $x(n)$ is linear therefore it admits $MA(\infty)$ representation with respect to a white process $\epsilon(n)$ with variance σ^2 and $p(\omega) = |H(\omega)|^2\sigma^2$. The following equality is the backbone of Hinich's bispectrum test:

$$\frac{p_2}{H(\omega_1)H(\omega_2)H^*(\omega_1+\omega_2)} = \frac{\mathbb{E}\{\epsilon(t)^3\}}{(\sigma^2)^{3/2}} = \mathcal{K}_3 \tag{135}$$

For a non-Gaussian linear process, the normalized bispectrum is constant, and zero for all frequency pairs if the process is linear Gaussian. A non-linear time series, on the other hand, exhibits a skewness function with bifrequency dependent magnitude. In Hinich's original test, the bispectrum is estimated by averaging the two-variable periodogram using a rectangular window:

$$\hat{B}_2(m,n) = M^{-2} \sum_{j=(m-1)M}^{mM-1} \sum_{k=(n-1)M}^{nM-1} F(j,k)$$

such that $F(j,k) = N^{-1}\hat{x}(\omega_j)\hat{x}(\omega_k)\hat{x}(\omega_{j+k})$. The arbitrariness of the choice of M was elegantly addressed in [124] by maximizing the test statistic over the feasible values of M. A kernel-smoothing version of Hinich's test was also proposed in [15]:

$$H_k = \frac{2\pi N}{\delta B^2} \sum_{i=1}^{k} \frac{|\hat{P}_2(\omega_{1,i}, \omega_{2,i})|^2}{\hat{P}_1(\omega_{1,i})\hat{P}_1(\omega_{2,i})\hat{P}_2(\omega_{1,i}+\omega_{2,i})} \tag{136}$$

where $\hat{p}$ and $\hat{p}_2$ are kernel-smoothed estimators of the spectral and bispectral density that are shown to converge faster, but the smoothing will inevitably induce bias and lead to higher false alarms.

$\{(\omega_{1,i}, \omega_{2,i})\}_{i=1\ldots k}$ are frequency pairs contained in the set W. B is a bandwidth parameter associated with the bispectrum estimation, and δ is a normalizing constant associated with $\hat{p}$. Under $\mathcal{H}_0$, H_k is distributed as χ^2_{2k}.

3.2.4 FOR d-DIMENSIONAL COLORED PROCESSES

GENERALIZATION OF HINICH'S BISPECTRUM

Hinich's bi-spectrum generalizes to multivariate series as proposed in [141]. The r^{th} order spectral vector for linear multivariate time-series is expressed as:

$$\mathbf{p}_r(\omega_1, \ldots, \omega_{r-1})(\mathbf{G}(\omega_1) \otimes \ldots \otimes \mathbf{G}(\omega_{r-1}))^{-1}\mathbf{p}^*_r(\omega_1, \ldots, \omega_{r-1}) = \boldsymbol{\kappa}_r^T\big(\underbrace{\boldsymbol{\Sigma} \otimes \ldots \otimes \boldsymbol{\Sigma}}_{r \text{ times}}\big)\boldsymbol{\kappa}_r \tag{137}$$

Wherein $\mathbf{x}(t) = \sum_{i=0}^{\infty} \mathbf{A}_i \boldsymbol{\epsilon}(t-i)$, $\mathbb{E}\{\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}(m)^T\} = \delta_n^m \boldsymbol{\Sigma}$ and $\mathbf{G}(\omega) = \mathbf{H}(\omega)\boldsymbol{\Sigma}\mathbf{H}^*(\omega)$, $\mathbf{H}(\omega) = \sum_{n=-\infty}^{\infty} \mathbf{A}_n \exp^{-j\omega n}$.

In particular for $r = 3$:

$$H_{ij} = \hat{\mathbf{p}}^*_2(\omega_i, \omega_j)(\mathbf{G}(\omega_i) \otimes \mathbf{G}(\omega_j) \otimes \mathbf{G}^*(\omega_i + \omega_j))^{-1}\hat{\mathbf{p}}_2(\omega_i, \omega_j)$$

can be approximated by a χ^2 distribution. This forms the basis for hypothesis testing of departures from multivariate Gaussianity and linearity. However, in practice, this procedure suffers from severe drawbacks. The spectral estimators require not only the careful choice of the smoothing window and its width, but also very large number of samples to converge (of the order of 100 of thousands) rendering their application for real-time responses very slow; moreover [18] have put Hinich's test under scrutiny and showed that it suffers from severe statistical problems. They propose the use of surrogate data to ensure the correct false alarm rate.

NON-LINEAR TIME-EMBEDDING

Back to a class of time-domain procedures where we consider a finite set N of observations $x(t)$. Authors in [106] apply a non-linear transformation to $x(t)$, for example if we are interested in third order moments, then a possible embedding is:

$$z(t) = [x(t), x(t)x(t+1),^2 x(t+1), x(t)^3, \ldots]^T \in \mathbb{R}^d$$

Their test statistic is based on the deviation of the sample $z(t)$ from its statistical mean:

Theorem 7 ([106]) *Let* $\hat{\mu}_i = \frac{1}{N}\sum_{t=1}^{N} z_i(t)$, *and* $\mu_i = \mathbb{E}\{z_i\}$ *for* $i \in \{1, \dots, d\}$. *For a strong-mixing 4.2.1 Gaussian process (under the null hypothesis):*

$$\sqrt{N}\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$$

Therefore the limiting distribution of the quadratic form L*:*

$$L = N\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)$$

is χ^2 *with* d *degrees of freedom.*

Our work can be seen as a subset of this class of procedures, where we focus on Mardia's kurtosis (therefore fourth order moments) in the general multivariate case $\mathbf{X} \in \mathbb{R}^{N\times d}$ and derive its limiting distribution. For that end, certain non-linear transformations of the type $x_a(n)^r x_b(j)^m x_c(i)$ appear in our computation up to order 16.

Chemistry between Gaussianity, linearity and stationarity tests

We mainly focused on tests aimed at characterizing Gaussianity, besides Hinich's procedure that tests both Gaussianity and linearity. As a matter of fact, there are not only procedures for testing linearity [15, 63, 69] but also for testing stationarity [4, 20] and serial dependence (whiteness) [38, 134]. Since these properties are closely related, as pointed out in the first chapter, the tests are naturally related in the sense that as explained in [33]:

- As pointed out by [54, 104], a test designed for white processes applied to dependent data overrejects the null hypothesis of normality.
- All the Gaussianity procedures assume stationarity. A non-stationary Gaussian process can therefore be misspecified as non-Gaussian.
- All Gaussian processes are linear, but there exists linear non-Gaussian processes. Non-linear processes are all non-Gaussian.

3.3 DISCUSSION AND CONCLUSIONS

On one hand, we show how testing normality for (scalar or multivariate) white processes has the lion's share of the proposed procedures. When the time-dependence enters the scene, test procedures become scarce for univariate time-series and even scarcer for multivariate time-series. Despite the efforts made on testing Gaussianity for univariate time-series, both in time and frequency domains, their shortcomings hinder their generalization to multivariate time-series.

Epps test's implementation is already difficult in the scalar case and requires many steps that cannot be efficiently translated in a real-time framework. The random projections method is powerful but it still depends on Epps test or Lobato-velasco for testing the univariate normality of the time-series. A possible solution in the same spirit as [118] would be to use auto-regressive seive bootstrap methods to approximate the limiting distribution of multivariate generalizations of Anderson-darling or Cramér-von-mises for time-series. For real-time procedures, this would require repeating the bootstrap steps multiple times.

In our work, the candidate that seems to fit best our computational constraints would be based on one of the descriptive measures: Skewness or Kurtosis. we carry on the work of Mardia, and the corrections of Gasser and Lobato-Velasco, by deriving the limiting distributions of Mardia's Kurtosis for multivariate time-series. If the calculations are rather involved, the computational burden is very low.

We are also aware of the shortcomings of this choice, and we do not claim that our proposed test fully characterizes normality against all alternatives i. e. as pointed out by Henze [67], the test procedure is only consistent for alternatives that satisfy $\beta_{2,d} \neq d(d+2)$. In the same spirit as [111] [96], the final testing procedure will be based on random bivariate projections, hence it is unlikely that all projections satisfy $\beta_{2,2} = 8$. Additionally, we use the a priori knowledge we have about the transients we want to detect: when a seismic signals arrives, the signals are very peaked in the beginning which translates as a heavy-tailed distribution.

4

DERIVING THE NORMALITY TEST

ABSTRACT

The following theoretical derivation of the test statistic were published in the European journal Signal processing. We reproduce in the following many of the equations and paragraphs that are present in [43]. Initially, Pierre & Laurent [31] were the first to formalize this problem and they have provided the limiting distribution of Mardia's test for a particular case of time-embedding time-series. They show that their test is applicable and exhibits a good power on both synthetic and real data. This has encouraged us to carry on their work, and derive the limiting distribution of Mardia's kurtosis in the general multivariate case.

Contents

The interest in techniques involving higher order statistics has grown considerably during the past decades [29, 32, 66, 112]. Actually, as we have seen in Chapter 1, first and second order statistics allow an exhaustive characterization of Gaussian processes and linear systems. Despite the practical importance of the Gaussian distribution, thanks to the central limit theorem, and the prevalence of linear dynamical systems in small fluctuations models, many situations do not resort to these assumptions. As a consequence, detecting departure from Gaussianity arose as a means to detect and characterize non linear behavior, detection of changes in dynamical regimes [11], etc. Higher-Order Statistics (HOS) were also shown to carry valuable information for blind identification problems, source separation and in measuring information theoretic quantities [32], to name a few applications.

The present growth of interest in sensor networks and our ability to simultaneously record time series representing the fluctuations of numerous physical quantities, naturally leads to consider d-dimensional processes. Surprisingly enough as we have seen in the previous chapter, normality tests for such d-dimensional stochastic processes were not so much investigated. Therefore, the purpose of this chapter is to propose a normality test that is simple to implement, even for colored (time correlated) d-dimensional processes, eventually at the expense of quite complicated and lengthy calculus to derive the exact form of the test. For this reason, we shall focus on the multivariate kurtosis proposed by Mardia in [98] for i.i.d. d-dimensional samples, and partially extended for colored samples in [31].

CONTRIBUTIONS. In a first contribution, we extend the results of [31] and the nature of the d-dimensional samples is no longer restricted to be obtained by time delay embedding. We give the exact formulas for the general case of a bivariate process, for instance a source observed by two sensors or the bivariate projection of the observations of a d-axis sensor. The latter results lead to the second contribution, generalizing the tests proposed in [96, 98] based on 1D projections and i.i.d samples, or in [111] for scalar n.i.d. samples. The benefits of using 2D is clear, as the resulting tests are subsequently shown to outperform 1D projection-based tests, via computer experiments. The importance of joint normality and the performance of our test is illustrated on *n.i.d.* copulas, i. e. with colored Gaussian marginals.

4.1 REINTRODUCING THE TEST STATISTIC

The population kurtosis measure of a d-variate process $\mathbf{X}$ as proposed by Mardia [98]:

$$\beta_{2,d} = \mathbb{E}\{(\mathbf{x}(n)^T \mathbf{S}^{-1} \mathbf{x}(n))^2\}. \tag{138}$$

Other definitions were suggested, for instance Mori et al. [105] defined the kurtosis as a $d \times d$ matrix:

$$\mathbf{B} = \mathbb{E}\{\mathbf{y}(n)\mathbf{y}(n)^T \mathbf{y}(n)\mathbf{y}(n)^T\} - (d+2)\mathbf{I}_d$$

wherein $\mathbf{y}(n) = \mathbf{S}^{-1/2}\mathbf{x}(n)$. Mardia's test can be seen as:[1]

$$\beta_{2,d} = \mathrm{Tr}(\mathbf{B} + (d+2)\mathbf{I}_d)$$

Malkovich and afifi [96] suggested using the definition of univariate definition of the kurtosis on linear combinations of $\mathbf{x}(n)$.

Lemma 4.1.1 *For normally distributed samples i. e.* $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{S})$*, one can easily show that* $\beta_d = d(d+2)$.

Proof.

$$\begin{aligned}
\beta_{2,d} &= \mathbb{E}\{\sum_{a,b} x_a(n)x_b(n)x_c(n)x_d(n)G_{ab}G_{cd}\} \\
&= \sum_{a,b,c,d} ([3]S_{ab}S_{cd})G_{ab}G_{cd} \\
&= \sum_{a,b} S_{ab}G_{ab} \sum_{c,d} S_{cd}G_{cd} + 2\sum_{abcd} S_{ac}G_{cd}S_{bd}G_{ba} \\
&= \mathrm{Tr}^2(\mathbf{SG}) + 2\mathrm{Tr}(\mathbf{SGSG}) = d^2 + 2d
\end{aligned}$$

□

PROPERTIES AND SHORTCOMINGS

Its sample counterpart for a sample of size N is:

$$B_d(N) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}(n)^T \mathbf{S}^{-1} \mathbf{x}(n))^2 \tag{139}$$

It is worth noticing that $\mathbf{S}$ being the exact covariance matrix, all random realizations involved in the latter equation are standardized

1 or alternatively as $\beta_{2,d} = \mathrm{Tr}(\mathbf{y}(n)\mathbf{y}(n)^T \otimes \mathbf{y}(n)\mathbf{y}(n)^T)$

(recall that we assume zero-mean processes). Thus, the advantage of this test variable is that it is invariant with respect to linear transformations, i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}$. In practice, the covariance matrix $\mathbf{S}$ is unknown and is replaced by its sample estimate, $\hat{\mathbf{S}}$, so that we end up with the following test variable:

$$\hat{B}_d(N) = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}(n)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(n)\right)^2 \tag{140}$$

with

$$\hat{\mathbf{S}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}(k)\mathbf{x}(k)^T. \tag{141}$$

One obvious advantage is that it characterizes d-variate random variables while being *scalar*. This is important from computational viewpoint. Perhaps the easiest way to give intuition to this measure is to see it as the the arithemtic mean of Mahalanobis distance $d_{nn} = \sqrt{\mathbf{x}(n)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(n)}$ raised to the fourth power i. e. $\hat{B}_d(N) = \frac{1}{N}\sum_{n=1}^{N} d_{nn}^4$.

A disadvantage was pointed out by Koziol [86] who noticed that in Mardia's definition only some entries of κ_4 are taken into account, namely:

$$\beta_{2,d} = \mathbb{E}\{\sum_{a=b} y_a^4 + \sum_{a\neq b} y_a^2 y_b^2\}$$

They suggested the alternative

$$B'_{2,d} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{y}(i)^T\mathbf{y}(j))^4$$

Hence, Mardia's Kurtosis can have the same numerical values for distributions with different shapes as stated and illustrated in [83]:

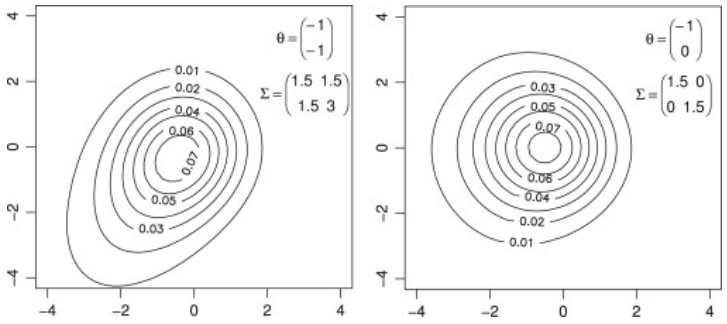


Figure 6: Two different bivariate asymetric Laplace distributions having the same value of Mardia's Kurtosis $\beta_{2,2} = 20$. Image taken from [83]

However, Mardia's measure is still suitable for hypothesis testing, since under the null hypothesis of a Gaussian distribution, (and in general for all elliptically symmetric distributions), r^{th} order moments which are odd are null and Mardia's kurtosis contains all the elements of cumulant vector κ_4.

Mardia has derived the limiting distribution of this measure for i.i.d d-dimensional processes, and we have stated the theorem in Chapter 3, in 3.2.2. Based on his findings, he proposes to test normality by testing $\beta_{2,2} = d(d+2)$.

Another formulation of Theorem 2 is:

$$z \stackrel{def}{=} (\hat{B}_d - d(d+2)\frac{N-1}{N+1})/\sqrt{8d(d+2)/N} \xrightarrow[N\to\infty]{} \mathcal{N}(0,1) \quad (142)$$

This shows that the quantities z and $\sqrt{N}(\hat{B}_d - \beta_{2,d})/\sqrt{8d(d+2)}$ are asymptotically equivalent as $N \to \infty$. The rejection of the null hypothesis is for large or small values of $\hat{B}_d$; it is interpreted as a departure from normality, in the sense that the multivariate kurtosis of $\mathbf{x}$, $\beta_{2,d}$, is sufficiently far from $d(d+2)$.

In practice, for large values of N, we can test the multinormality of $\mathbf{x}$ by comparing z to the critical values $\pm 1,96$ of a standard normal (for a test level of 5%). If $|z| \leqslant 1,96$ then the gap is not significant and we cannot reject the assumption of normality.

The objective of this chapter is to derive the limiting distribution of Mardia's measure, for *n.i.d* d-dimensional processes. Since this involves heavy calculations, we need to introduce some preliminary tools.

4.2 ASSUMPTIONS AND LEMMAS

In this section, partial useful results are established. Each is associated with a lemma, and represents a step towards the derivation of the exact expression of the statistics of $\widehat{B}_d(N)$ defined in (140) for multivariate *colored* processes:

- Lemma 4.2.1 proves that $\boldsymbol{\Delta} = \hat{\mathbf{S}} - \mathbf{S}$ varies as $O(1/\sqrt{N})$.
- Lemma 4.2.2 uses the preceding result in order to express the sample precision matrix $\hat{\mathbf{G}} = (\mathbf{S} + \boldsymbol{\Delta})^{-1}$ as a function of the exact precision matrix $\mathbf{G}$ and of the approximation matrix $\boldsymbol{\Delta}$, up to order $O(\|\boldsymbol{\Delta}\|^3)$.
- Finally, lemma 4.2.3 allows to derive the approximate expression of $\hat{B}_p(N)$ in $o(1/N)$.

From now on, we assume a mixing condition upon $\mathbf{x}(n)$, necessary to relax the i.i.d. property while maintaining convergence of various terms. Recall that we have met this property in Chapter 1 and established it link with ergodicity. We now make use of it in order to derive the limiting distribution of our test statistic.

Assumption 4.2.1 (α-mixing) *Under $\mathcal{H}_0$, $\mathbf{x}(n)$ is a stationary Gaussian linear process and $|S(\tau)| \sim O(\rho^{|\tau|})$ with $\rho \in (0,1)$ for large $|\tau|$. Hence the sequence $\mathbf{x}(1),\dots,\mathbf{x}(N)$ is α-mixing with $\alpha_n = \rho^n$ (also called strongly mixing since $\alpha_n \to 0$). One consequence is that $\sum_{\tau=0}^{\infty} |S_{ab}(\tau)|^2$ converges to a finite limit Ω_{ab}, $\forall (a,b) \in \{1,\dots,p\}^2$, where S_{ab} denote the entries of matrix $\mathbf{S}$.*

Remark 4.2.1 *Assumption 4.2.1 is not restrictive in our framework. In fact, theorems 1 and 2 in [84] imply that a Gaussian process is strongly mixing if and only if the maximal correlation $S_{ab}(\tau) \to 0$ when $\tau \to \infty$. This condition is satisfied here since $\mathbf{x}$ can be represented as an autoregressive process $AR(m)$. Note that the order m may be very large but will in practice remain finite for finite time series modeling. Hence, the correlation function $|S(\tau)|$ decays exponentially as $\rho^{|\tau|}$, where $0 < \rho < 1$ is the modulus of the largest pole of the linear AR-filter modeling the series.*

LEMMAS

The estimated multivariate kurtosis (140) is a rational function of degree 4 in $\mathbf{x}$. Since we wish to calculate its asymptotic first and second order moments, when N tends to infinity, we may expand this rational function about its mean. The first step is to expand the estimated covariance $\hat{\mathbf{S}}$. Let $\hat{\mathbf{S}} = \mathbf{S} + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ is small compared to $\mathbf{S}$; we have the following lemmas :

the proofs are deferred to Appendix A.1.1, for sake of readability

Lemma 4.2.1 *The entries of matrix $\boldsymbol{\Delta}$ are of order $O(1/\sqrt{N})$.*

Lemma 4.2.2 *The inverse $\hat{\mathbf{G}}$ of $\hat{\mathbf{S}}$ can be approximated by*

$$\hat{\mathbf{G}} = \mathbf{G} - \mathbf{G}\boldsymbol{\Delta}\mathbf{G} + \mathbf{G}\boldsymbol{\Delta}\mathbf{G}\boldsymbol{\Delta}\mathbf{G} + o(1/N). \tag{143}$$

In order to express $\hat{\mathbf{G}}$ as a function of $\hat{\mathbf{S}}$, we replace $\boldsymbol{\Delta}$ by $\hat{\mathbf{S}} - \mathbf{S}$ in (143), and obtain:

$$\hat{\mathbf{G}} = 3\mathbf{G} - 3\mathbf{G}\hat{\mathbf{S}}\mathbf{G} + \mathbf{G}\hat{\mathbf{S}}\mathbf{G}\hat{\mathbf{S}}\mathbf{G} + o(1/N). \tag{144}$$

With this approximation, $\hat{\mathbf{G}}$ is now a polynomial function of $\hat{\mathbf{S}}$ of degree 2, and hence of degree 4 in $\mathbf{x}$. We shall show that the mean of $\hat{B}_p(N)$ involves moments of $\mathbf{x}$ up to order 8, whereas its variance involves moments up to order 16.

Lemma 4.2.3 *Denote* $A_{ij} = \mathbf{x}(i)^T \mathbf{S}^{-1} \mathbf{x}(j)$. *Then:*

$$\hat{B}_d(N) = \frac{6}{N}\sum_{n=1}^{N} A_{nn}^2 - \frac{8}{N^2}\sum_{n=1}^{N} A_{nn}\sum_{i=1}^{N} A_{ni}^2 + \frac{1}{N^3}\sum_{n=1}^{N}\Big(\sum_{i=1}^{N} A_{ni}^2\Big)\Big(\sum_{j=1}^{N} A_{nj}^2\Big) + \frac{2}{N^3}\sum_{n=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} A_{nn}A_{nj}A_{jk}A_{kn} + o(1/N) \tag{145}$$

The objective of this lemma is to derive an expression of $\hat{B}_p$ involving the exact covariance $\mathbf{S}$ and not its sample counterpart $\hat{\mathbf{S}}$.

Proof. First inject (143) in the expression, and keep terms up to order $O(\|\boldsymbol{\Delta}\|^2)$; this yields:

$$\hat{B}_d(N) = \frac{1}{N}\sum_{n}\Big[A_{nn}^2 - 2A_{nn}\,\mathbf{x}(n)^T\mathbf{G}\boldsymbol{\Delta}\mathbf{G}\mathbf{x}(n) + \big(\mathbf{x}(n)^T\mathbf{G}\boldsymbol{\Delta}\mathbf{G}\mathbf{x}(n)\big)^2 + 2A_{nn}\,\mathbf{x}(n)^T\mathbf{G}\boldsymbol{\Delta}\mathbf{G}\boldsymbol{\Delta}\mathbf{G}\mathbf{x}(n)\Big] + o(\|\boldsymbol{\Delta}\|^2).$$

Then replace $\boldsymbol{\Delta}$ by $\hat{\mathbf{S}} - \mathbf{S}$. This leads to

$$\hat{B}_d(N) = \frac{1}{N}\sum_{n}\Big[6A_{nn}^2 - 8A_{nn}\big(\mathbf{x}(n)^T\mathbf{G}\hat{\mathbf{S}}\mathbf{G}\mathbf{x}(n)\big) + \big(\mathbf{x}(n)^T\mathbf{G}\hat{\mathbf{S}}\mathbf{G}\mathbf{x}(n)\big)^2 + 2A_{nn}\big(\mathbf{x}(n)^T\mathbf{G}\hat{\mathbf{S}}\mathbf{G}\hat{\mathbf{S}}\mathbf{G}\mathbf{x}(n)\big)\Big] + o(\|\boldsymbol{\Delta}\|^2).$$

Equation (145) is eventually obtained after replacing $\hat{\mathbf{S}}$ by $\frac{1}{N}\sum_k \mathbf{x}(k)\mathbf{x}(k)^T$ and all terms of the form $\mathbf{x}(q)^T\mathbf{G}\mathbf{x}(r)$ by A_{qr}. □

4.2.1 CALCULUS ISSUES

When computing the mean and variance of $\hat{B}_d(N)$ given in (145), higher order moments of the multivariate random variable $\mathbf{x}$ will arise. Under the normal (null) hypothesis, these moments are expressed as functions of second order moments only. To keep notations reasonably concise, it is proposed to use McCullagh's bracket notation [101], briefly reminded in Appendix A.1.2. Furthermore, for all moments of order higher than d, some components appear multiple times; counting the number of identical terms in the expansion of the higher moments is a tedious task. All the moment expansions that are necessary for the derivations presented in this paper are developed in Appendix A.1.4.

In order to keep notations as explicit and concise as possible, while keeping explicit the role of both coordinate (or space) indices and time

indices, let the moments of $\mathbf{x}(t)$, whose d components are $x_a(t)$, $1 \leqslant a \leqslant d$ be noted

$$\mu_{ab}^{tu} = \mathbb{E}\{x_a(t)x_b(u)\}, \quad \mu_{abc}^{tuv} = \mathbb{E}\{x_a(t)x_b(u)x_c(v)\} \tag{146}$$

and so forth for higher orders. It shall be emphasized that different time and coordinate indices appear here as the components are assumed to be colored (time correlated) and dependent to each others (spatially correlated).

Computation of the mean and variance of $\hat{B}_p$ defined by equation (145) involves the computation of moments of order noted $2L$ whose generic expression is

$$\mathbb{E}\{\prod_{l=1}^{L} A_{\alpha^l\beta^l}\} = \sum_{r_1\ldots r_L, c_1\ldots c_L=1}^{d} \left(\prod_{i=1}^{L} G_{r_i,c_i}\right) \mu_{r_1\ldots r_L c_1\ldots c_L}^{\alpha_1\ldots\alpha_L\beta_1\ldots\beta_L} \tag{147}$$

In the above equation, the $2L$-order moment $\mu_{r_1\ldots r_L c_1\ldots c_L}^{\alpha_1\ldots\alpha_L\beta_1\ldots\beta_L}$ has superscripts indicating the time indices involved, whereas the subscripts indicate the coordinate (or space) indices.

While being general, the above formulation may take simpler, or more explicit forms in practice. The detailed methodology for computing the expressions of the mean and variance of $\hat{B}_d$ as functions of second order moments is deferred to Appendix A.1.3. The resulting expressions of Mardia's statistics are given and discussed in the sections to come.

4.3 EXPRESSION OF THE MEAN OF $\widehat{B}_d(N)$

According to Equation (145), we have four types of terms. The goal of this section is to provide the expectation of each of these terms. In the propositions below, all terms are developed as being sums and products of second order moments, as it is reminded that under $\mathcal{H}_0$ the process is Gaussian. Notice also that under the latter assumption, all higher-order moments of any order are finite. For sake of simplicity, Landau's approximation order $O(h(n))$ is omitted in most equations.

Lemma 4.3.1 *With the definition of A_{ij} given in Lemma 4.2.3, we have:*

$$\mathbb{E}\{A_{nn}^2\} = \sum_{a,b,c,d=1} G_{ab}G_{cd}\,\mu_{abcd}^{nnnn} \tag{148}$$

$$\mathbb{E}\{A_{nn}A_{ni}^2\} = \sum_{a,b,c,d=1}\sum_{e,f=1} G_{ab}G_{cd}G_{ef}\,\mu_{abcedf}^{nnnnii} \tag{149}$$

$$\mathbb{E}\{A_{ni}^2A_{nj}^2\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1} G_{ab}G_{cd}G_{ef}G_{gh}\,\mu_{acegbdfh}^{nnnniijj} \tag{150}$$

$$\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1} G_{ab}G_{cd}G_{ef}G_{gh}\,\mu_{abchdefg}^{nnnnjjkk} \tag{151}$$

Proposition 4.3.1 *Using expressions of moments given in Appendix A.1.4,the expectations of the four terms defined in Lemma 4.3.1 take the form below*

$$\mathbb{E}\{A_{nn}^2\} = \sum_{k\ell qr=1} G_{k\ell}G_{rq}\Big\{[3]\mu_{k\ell}^{nn}\mu_{qr}^{nn}\Big\}$$

$$\mathbb{E}\{A_{nn}A_{ni}^2\} = \sum_{k\ell qrst=1} G_{k\ell}G_{qr}G_{st}\Big\{[12]\mu_{kr}^{ni}\mu_{\ell t}^{ni}\mu_{qs}^{nn} + [3]\mu_{k\ell}^{nn}\mu_{qs}^{nn}\mu_{rt}^{ii}\Big\}$$

$$\begin{aligned}\mathbb{E}\{A_{ni}^2A_{nj}^2\} = \sum_{k,\ell,q,r}\sum_{s,t,u,v} G_{k\ell}G_{qr}G_{st}G_{uv}\Big\{&[3]\mu_{kq}^{nn}\mu_{su}^{nn}\mu_{\ell r}^{ii}\mu_{tv}^{jj}\\ &+[6]\mu_{kq}^{nn}\mu_{su}^{nn}\mu_{\ell t}^{ij}\mu_{rv}^{ij} + [12]\mu_{kq}^{nn}\mu_{s\ell}^{ni}\mu_{ur}^{ni}\mu_{tv}^{jj} + [24]\mu_{kt}^{nj}\mu_{qv}^{nj}\mu_{\ell s}^{in}\mu_{ur}^{ni}\\ &+[48]\mu_{k\ell}^{ni}\mu_{rt}^{ij}\mu_{qv}^{nj}\mu_{su}^{nn} + [12]\mu_{kq}^{nn}\mu_{ts}^{jn}\mu_{uv}^{nj}\mu_{r\ell}^{ii}\Big\}\end{aligned}$$

$$\begin{aligned}\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}\} = \sum_{m,\ell,q,r}\sum_{s,t,u,v} G_{m\ell}G_{qr}G_{st}G_{uv}\Big\{&[3]\mu_{m\ell}^{nn}\mu_{qv}^{nn}\mu_{sr}^{jj}\mu_{tu}^{kk}\\ &+[6]\mu_{m\ell}^{nn}\mu_{qv}^{nn}\mu_{rt}^{jk}\mu_{su}^{jk} + [12]\mu_{m\ell}^{nn}\mu_{qr}^{nj}\mu_{vs}^{nj}\mu_{tu}^{kk}\\ &+[24]\mu_{mv}^{nk}\mu_{\ell u}^{nk}\mu_{qr}^{nj}\mu_{vs}^{nj} + [48]\mu_{mr}^{nj}\mu_{st}^{jk}\mu_{\ell u}^{nk}\mu_{qv}^{nn}\\ &+[12]\mu_{k\ell}^{nn}\mu_{qt}^{nk}\mu_{vu}^{nk}\mu_{rs}^{jj}\Big\}\end{aligned}$$

The mean of $\hat{B}_d(N)$ then follows from (145).

4.4 EXPRESSION OF THE VARIANCE OF $\widehat{B}_d(N)$

From Lemma 4.2.3, we can also state what moments of A_{ij} will be required in the expression of the variance of $B_p(N)$.

Lemma 4.4.1 *By raising (145) to the second power and using the definition of A_{ij} given in Lemma 4.2.3, we can check that the following moments are required:*

$$\mathbb{E}\{A_{nn}^2 A_{ii}^2\} = \sum_{a,b,c,d,e,f,g,h=1} G_{ab}G_{cd}G_{ef}G_{gh}\,\mu_{abcdefgh}^{nnnniiii}$$

$$\mathbb{E}\{A_{nn}^2 A_{ij}^2 A_{ii}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell=1} G_{ab}G_{cd}G_{ef}G_{gh}\,G_{m\ell}\,\mu_{abcdegm\ell fh}^{nnnniiiij}$$

$$\mathbb{E}\{A_{nn}A_{kk}A_{ni}^2A_{kj}^2\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1} G_{ab}G_{cd}G_{ef}G_{gh}G_{m\ell}\,G_{qr}\,\mu_{abegcdmqfh\ell r}^{nnnnkkkkiijj}$$

$$\mathbb{E}\{A_{kk}^2A_{ni}^2A_{nj}^2\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1} G_{ab}G_{cd}G_{ef}G_{gh}G_{m\ell}\,G_{qr}\,\mu_{abcdegmqfh\ell r}^{kkknnnniijj}$$

$$\mathbb{E}\{A_{nn}^2A_{ii}A_{ij}A_{jk}A_{ki}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1} G_{ab}G_{cd}G_{ef}G_{gh}G_{m\ell}\,G_{qr}\,\mu_{abcdefgrhm\ell q}^{nnnniiiijjkk}$$

$$\mathbb{E}\{A_{ni}^2A_{nj}^2A_{kt}^2A_{kk}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1}\sum_{s,u=1} G_{ab}G_{cd}G_{ef}G_{gh}\,G_{m\ell}G_{qr}G_{su}\,\mu_{acegmqsubdfh\ell r}^{nnnnkkkkiijjtt}$$

$$\mathbb{E}\{A_{ii}A_{it}^2A_{nn}A_{nj}A_{jk}A_{kn}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1}\sum_{s,u=1} G_{ab}G_{cd}G_{ef}G_{gh}\,G_{m\ell}G_{qr}G_{su}\,\mu_{abceghmu\ell qrsdf}^{iiiinnnnjjkktt}$$

$$\mathbb{E}\{A_{ni}^2A_{kt}^2A_{nj}^2A_{ku}^2\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1}\sum_{s,v,w,z=1} G_{ab}G_{cd}G_{ef}\,G_{gh}G_{m\ell}G_{qr}G_{sv}G_{wz}\,\mu_{acmqegswbd\ell rfhvz}^{nnnnkkkkiijjttuu}$$

$$\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}A_{ii}A_{it}A_{tu}A_{ui}\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1}\sum_{s,v,w,z=1} G_{ab}G_{cd}G_{ef}\,G_{gh}G_{m\ell}G_{qr}G_{sv}G_{wz}\,\mu_{abchm\ell qzdefgrsvw}^{nnnniiiijjkkttuu}$$

$$\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}A_{it}^2A_{iu}^2\} = \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}\sum_{m,\ell,q,r=1}\sum_{s,v,w,z=1} G_{ab}G_{cd}G_{ef}\,G_{gh}G_{m\ell}G_{qr}G_{sv}G_{wz}\,\mu_{abchmqswdefg\ell rvz}^{nnnniiiijjkkttuu}$$

Then, as in Proposition 4.3.1, by using the results of Appendix A.1.4, the moments μ_*^* could be in turn expressed as a function of second order moments. For readability, we do not substitute here these values.

Proposition 4.4.1

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\{\hat{B}_d\} \;=\; & \frac{36}{N^2}\sum_{n}\sum_{i}\mathbb{E}\{A_{nn}^2A_{ii}^2\}-\frac{96}{N^3}\sum_{j}\sum_{n,i}\mathbb{E}\{A_{nn}^2A_{ij}^2A_{ii}\}\\
&+\frac{64}{N^4}\sum_{n,i}\sum_{j,k}\mathbb{E}\{A_{nn}A_{kk}A_{ni}^2A_{kj}^2\}+\frac{12}{N^4}\sum_{n,i,j,k}\mathbb{E}\{A_{kk}^2A_{ni}^2A_{nj}^2\}\\
&+\frac{24}{N^4}\sum_{n,i,j,k}\mathbb{E}\{A_{nn}^2A_{ii}A_{ij}A_{jk}A_{ki}\}-\frac{16}{N^5}\sum_{n,i}\sum_{j,k,t}\mathbb{E}\{A_{ni}^2A_{nj}^2A_{kt}^2A_{kk}\}\\
&-\frac{32}{N^5}\sum_{i,t}\sum_{n,j,k}\mathbb{E}\{A_{ii}A_{it}^2A_{nn}A_{nj}A_{jk}A_{kn}\}\\
&+\frac{1}{N^6}\sum_{n,i,j}\sum_{k,t,u}\mathbb{E}\{A_{ni}^2A_{kt}^2A_{nj}^2A_{ku}^2\}\\
&+\frac{4}{N^6}\sum_{n,j,k}\sum_{i,t,u}\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}A_{ii}A_{it}A_{tu}A_{ui}\}\\
&+\frac{4}{N^6}\sum_{n,j,k}\sum_{i,t,u}\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}A_{it}^2A_{iu}^2\}-(\mathbb{E}\{\hat{B}_d\})^2 \qquad (152)
\end{aligned}$$

4.5 MAIN RESULT

4.5.1 ASYMPTOTIC DISTRIBUTION OF $\hat{B}_d(N)$

(i) we first prove that $\hat{B}_d(N)$ converges towards $B_d(N)$ as N tends to infinity. Expressions (153), (155) imply

$$\mathbb{E}\{\hat{B}_d - B_d\} = O(\frac{1}{N})$$

since $S_{ab}^2(\tau)$ is bounded (cf. Assumption 4.2.1), and since $q_1(\tau)$, a linear combination of $S_{ab}(\tau)S_{cd}(\tau)$, $a, b, c, d \in \{1, \ldots, d\}$, is bounded. Next, expressions (154), (156) ensure that $\mathbb{V}\mathrm{ar}\{\hat{B}_d\} \underset{N\to+\infty}{\longrightarrow} 0$; in fact, $Q_2(\tau)$ involve fourth order moments, which under $\mathcal{H}_0$ can be expanded as a function of cross-correlations of the type of $S_{ab}(\tau)$. By similar arguments as before, the convergence in the mean square sense $\hat{B}_d \underset{N\to+\infty}{\longrightarrow} B_d$ is eventually obtained.

It remains thus to establish the convergence of B_d (instead of $\hat{B}_d$) towards a normal distribution. With this goal, introduce the centered random variable

$$w(n) \overset{\text{def}}{=} (\mathbf{x}(n)^T\mathbf{S}^{-1}\mathbf{x}(n))^2 - \mathbb{E}\{(\mathbf{x}(n)^T\mathbf{S}^{-1}\mathbf{x}(n))^2\} \overset{\text{def}}{=} A_{nn}^2 - \mathbb{E}\{A_{nn}^2\}.$$

The remainder of the proof goes in two stages: (ii) prove that $w(n)$ is mixing as soon as $\mathbf{x}(n)$ is mixing, and (iii) prove that $B_d(N)$ is asymptotically normal.

(ii) For now, we assume for simplicity that $x(n) \sim \mathcal{N}(0,1)$ is a scalar process. This simplifies significantly the writing of the proof, but does not restrict its generality since the steps remain the same in the multivariate case – up to more cumbersome notation.

Its probability density function is denoted $P_{x(n)}$. Recall that x is mixing (Assumption 4.2.1) thus the joint density function $P_{x(n),x(i)} \approx P_{x(n)} \times P_{x(i)}$ for sufficiently large time intervals. Now, consider the new variable $w(n) = x^4(n)$ whose probability density function can be expressed as:

$$P_{w(n)} = \frac{1}{2\sqrt{2\pi}w(n)^{\frac{3}{4}}} \exp\left(-\frac{1}{2}\sqrt{w}\right)\mathbb{1}_{w(n)>0}$$

$$\begin{aligned} P_{w(n),w(i)} &= P_{x(n),x(i)} \times (\det J)^{-1} \\ &= \frac{1}{16w(n)^{\frac{3}{4}}w(p)^{\frac{3}{4}}\sqrt{2\pi}\det S^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x(n)^2 + x(p)^2\} \\ &\quad + \mathbb{E}\{x(n)x(p)\}))\mathbb{1}_{w(n)>0}\mathbb{1}_{w(p)>0} \\ &\approx P_{z(n)}P_{z(i)}\frac{1}{\sqrt{1-\rho^{2|n-i|}}} \exp(-x(n)x(i)\rho^{|n-i|}) \end{aligned}$$

where $\det J$ denotes the determinant of the Jacobian of the transformation $x(n), x(p) \to w(n), w(p)$. S is the covariance matrix of the jointly Gaussian process $(x(n), x(p))$. We recall that, from Assumption 4.2.1, $|\mathbb{E}\{x(n)x(p)\}| \overset{\text{def}}{=} |S(n-p)| = O(\rho^{|n-p|})$.

Hence, x being α-mixing implies that $w(n)$ is α-mixing.

(iii) We have now established that the sequence $w(1), \ldots, w(N)$ is stationary and strongly mixing. Moreover:

$$N^{-1}\mathbb{V}\mathrm{ar}\{\sum_{n=1}^{N} w(n)\} = \mathbb{E}\{w(0)^2\} + 2\sum_{\tau=1}^{N-1}\mathbb{E}\{w(1)w(\tau+1)\}$$

Let $n, i \in \{1, \ldots, N\}$, we have $\mathbb{E}\{A_{nn}\} = d$ since A_{nn} is χ^2_d-distributed and:

$$\begin{aligned} \mathbb{E}\{A_{nn}^2\} &= d(d+2) \overset{\text{def}}{=} c \\ \mathbb{E}\{w(n)w(i)\} &= \mathbb{E}\{A_{nn}^2A_{ii}^2\} - c^2 \\ &= \sum_{a,b,c,d=1}\sum_{e,f,g,h=1}^{d} G_{ab}G_{cd}G_{ef}G_{gh}\mu^{nnnniiii}_{abcdefgh} - c^2 \end{aligned}$$

$$\mathbb{E}\{w(n)w(i)\} = \sum_{a,b,c,d=1}^{d} \sum_{e,f,g,h=1} G_{ab}G_{cd}G_{ef}G_{gh}\Big([9]\mu_{ab}^{nn}\mu_{cd}^{nn}\mu_{ef}^{ii}\mu_{gh}^{ii} + [72]\mu_{ab}^{nn}\mu_{ce}^{ni}\mu_{df}^{ni}\mu_{gh}^{ii} + [24]\mu_{ae}^{ni}\mu_{bf}^{ni}\mu_{cg}^{ni}\mu_{dh}^{ni}\Big) - c^2$$

From Assumption 4.2.1, we have under $\mathcal{H}_0$ that $\mu_{ab}^{ni} \sim \rho^{|n-i|}$, where ρ is the decay rate of the covariance function.

In the sum above, addition of all terms involving factors of type μ_{ab}^{nn} only will be cancelled by the term c^2. Thus, $\mathbb{E}\{w(n)w(i)\}$ is bounded when $n, i \to \infty$ and converges absolutely.

To sum up, the sequence $w(1), \ldots, w(N)$ is stationary and strongly mixing. Moreover, $N^{-1}\mathbb{V}ar\{\sum_{n=1}^{N} w(n)\} \to \sigma^2$ when $N \to \infty$. The mixing rate decays exponentially, and following a similar reasoning as before (decomposition of higher order moments in terms of cross-covariances of $\mathbf{x}$ under $\mathcal{H}_0$) $\mathbb{E}\{w^{12}\} < \infty$.

All necessary conditions of Theorem [17, Thm 27.4] are verified, so that we can deduce that $B_d = \sum_{n=1}^{N} w(n)$ converges to a Normal distribution. Hence, it remains only to calculate the expressions of the mean and variance of $\hat{B}_d(N)$.

Empirical investigation of the limiting distribution of $\hat{B}_2$

Additionally, we conduct Monte Carlo simulation to verify empirically the limiting distribution of $\hat{B}_2$ on a colored process obtained by auto-regressive filtering. The kernel estimate of the probability density function of $\hat{B}_2$ is shown in Figure 7.

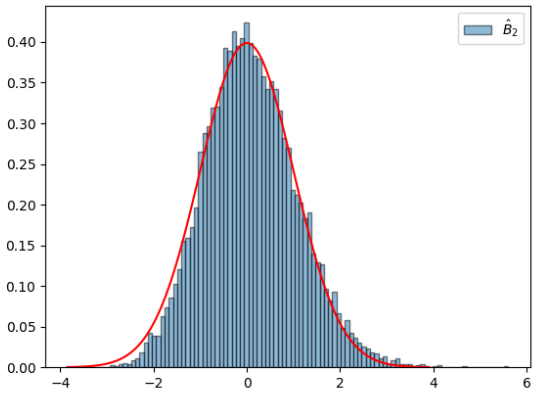


Figure 7: Histogram of 10000 realizations of $\hat{B}_2$ estimated on a VAR(1) time-series with $N = 1000$ observations.

4.5.2 MEAN AND VARIANCE OF $\hat{B}_1(N)$ IN THE SCALAR CASE ($d = 1$)

The complicated expressions obtained in the previous sections simplify drastically in the scalar case, and we get this expression for the mean:

$$\boxed{\mathbb{E}\{\hat{B}_1\} = 3 - \frac{6}{N} - \frac{12}{N^2}\sum_{\tau=1}^{N-1}(N-\tau)\,\frac{S(\tau)^2}{S^2} + o(\frac{1}{N})} \tag{153}$$

$$\boxed{Var\{\hat{B}_1\} = \frac{24}{N}\Big[1 + \frac{2}{N}\sum_{\tau=1}^{N-1}(N-\tau)\frac{S(\tau)^4}{S^4}\Big] + o(\frac{1}{N})} \tag{154}$$

Here are some intermediate steps for deriving $\mathbb{E}\{\hat{B}_1\}$:

$$\begin{aligned}
\mathbb{E}\{A_{nn}^2\} &= 3 \\
\mathbb{E}\{A_{nn}A_{ni}^2\} &= 3 + 12\frac{S(n-i)^2}{S^2} \\
\mathbb{E}\{A_{ni}^2A_{nj}^2\} &= 3 + 6\frac{S(i-j)^2}{S^2} + 12\frac{S(n-i)^2}{S^2} + 12\frac{S(n-j)^2}{S^2} \\
&\quad + 24\frac{S(n-i)^2S(n-j)^2}{S^4} + 48\frac{S(n-i)S(n-j)S(i-j)}{S^3} \\
\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}\} &= 3 + 6\frac{S(j-k)^2}{S^2} + 12\frac{S(n-k)^2}{S^2} + 12\frac{S(n-j)^2}{S^2} \\
&\quad + 24\frac{S(n-k)^2S(n-j)^2}{S^4} + 48\frac{S(n-k)S(j-k)S(n-j)}{S^3}
\end{aligned}$$

The exact computation of $\mathbb{E}\{\hat{B}_1\}$ yields the following result:

$$\begin{aligned}
\mathbb{E}\{\hat{B}_1\} &= 3 - \frac{6}{N^2}\sum_{n,i}\frac{S(n-i)^2}{S^2} + \frac{72}{N^3}\sum_{n,i,j}\frac{S(n-j)^2S(n-i)^2}{S^4} \\
&\quad + \frac{144}{N^3}\sum_{n,i,j}\frac{S(n-i)S(i-j)S(n-j)}{S^3}
\end{aligned}$$

Based on the results in [35, p. 346-347], it can be shown that $\frac{1}{N^3}\sum_{n,i,j}\frac{S(n-j)^2S(n-i)^2}{S^4}$ and $\frac{1}{N^3}\sum_{n,i,j}\frac{S(n-i)S(i-j)S(n-j)}{S^3}$ will contribute quantities of order lower than N^{-1}. Thus, we obtain the result in Equation 153 for the mean:

The second-order moment of $\hat{B}_1$ reads:

$$\mathbb{E}\{\hat{B}_1^2\} = 9 - \frac{36}{N^2}\sum_{n,k}\frac{S(n-k)^2}{S^2} + \frac{24}{N^2}\sum_{n,k}\frac{S(n-k)^4}{S^4} + o(\frac{1}{N})$$

and by raising to the second order the result of Equation 153, we have $\mathbb{V}ar\{\hat{B}_1\} = \mathbb{E}\{\hat{B}_1^2\} - 9 + \frac{36}{N^2}\sum_{n,k}\frac{S(n-k)^2}{S^2} + o(\frac{1}{N})$. In particular in the i.i.d. case, $S(\tau) = 0$ for $\tau \neq 0$, and we get the well-known result:

$$\mathbb{E}\{\hat{B}_1\} \approx 3 - \frac{6}{N}, \quad \text{and} \quad \mathbb{V}ar\{\hat{B}_1\} \approx \frac{24}{N}.$$

The expressions of mean and variance above are identical to those given in Theorem 2, the difference being that here the ratio $\frac{N-1}{N+1}$ is replaced by its approximation of order N^{-1}, i.e. $\frac{N-1}{N+1} = 1 - \frac{2}{N} + o(1/N)$.

The intermediate steps for calculating the expression of the variance (Equation 154) can be found in [45].

4.5.3 MEAN AND VARIANCE OF $\widehat{B}_2(N)$ IN THE BIVARIATE CASE $(d = 2)$

In the bivariate case, expressions become immediately more complicated as more moments are involved, but following the same pattern as before, we can still write them explicitly, as reported below. We remind that $\mu_{ab}^{ij} = S_{ab}(i-j)$.

$$\boxed{\mathbb{E}\{\hat{B}_2\} = 8 - \frac{16}{N} - \frac{4}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)Q_1(\tau)}{(S_{11}S_{22} - S_{12}^2)^2} + o(\frac{1}{N})} \tag{155}$$

$$\boxed{\mathbb{V}ar\{\hat{B}_2\} = \frac{64}{N} + \frac{16}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)Q_2(\tau)}{(S_{11}S_{22} - S_{12}^2)^4} + o(\frac{1}{N})} \tag{156}$$

where

$$\begin{aligned} Q_1(\tau) = {} & S_{11}S_{22}\Big[(S_{12}(\tau) + S_{21}(\tau))^2 - 4S_{11}(\tau)S_{22}(\tau)\Big] + S_{12}^2 \\ & \Big[2(S_{12}(\tau) + S_{21}(\tau))^2 + 4S_{22}(\tau)S_{11}(\tau)\Big] - 6S_{22}S_{12}\Big(S_{11}(\tau)(S_{12}(\tau) + S_{21}(\tau))\Big) \\ & -6S_{11}S_{12}\Big(S_{22}(\tau)(S_{12}(\tau) + S_{21}(\tau)\Big) + 6S_{11}^2S_{22}^2(\tau) + 6S_{22}^2S_{11}^2(\tau) \end{aligned}$$

and

$$
\begin{aligned}
Q_2(\tau) \;=\; & \Big[2S_{11}^2(\tau)S_{22}^2(\tau) - 16S_{11}(\tau)S_{22}(\tau)S_{12}(\tau)S_{21}(\tau) \\
& +3(S_{21}^2(\tau)+S_{12}^2(\tau))^2 + 12S_{11}(\tau)S_{22}(\tau)(S_{12}(\tau)+S_{21}(\tau))^2 \\
& -4S_{12}^2(\tau)S_{21}^2(\tau)\Big]S_{11}^2S_{22}^2 \\
& +2S_{11}^2S_{12}^2\Big[8S_{11}(\tau)S_{22}(\tau) + 3(5S_{11}(\tau)S_{22}(\tau) + S_{21}(\tau)S_{12}(\tau)) \\
& (S_{21}(\tau)+S_{12}(\tau))^2 - 4S_{21}(\tau)S_{12}(\tau)\big(S_{22}^2(\tau)+S_{21}(\tau)S_{12}(\tau)\big)\Big] \\
& +2S_{22}^2S_{12}^2\Big[8S_{22}(\tau)S_{11}(\tau) + 3(5S_{11}(\tau)S_{22}(\tau)) + S_{21}(\tau)S_{12}(\tau) \\
& (S_{21}(\tau)+S_{12}(\tau))^2 - 4S_{21}(\tau)S_{12}(\tau)\big(S_{11}^2(\tau)+S_{21}(\tau)S_{12}(\tau)\big)\Big] \\
& +3S_{11}^4S_{22}^4(\tau) + 3S_{22}^4S_{11}^4(\tau) \\
& +8S_{12}^4\Big[S_{11}^2(\tau)S_{22}^2(\tau) + 4S_{11}(\tau)S_{22}(\tau)S_{12}(\tau)S_{12}(\tau) + S_{21}^2(\tau)S_{12}^2(\tau)\Big] \\
& -12S_{11}S_{12}\,S_{22}(\tau)(S_{12}(\tau)+S_{21}(\tau)\Big[(2S_{11}(\tau)S_{22}(\tau) + S_{21}^2(\tau) + S_{12}^2(\tau)) \\
& S_{11}S_{22} + 2(S_{11}(\tau)S_{22}(\tau) + S_{12}(\tau)S_{21}(\tau))S_{12}^2\Big] \\
& -12S_{22}S_{12}\,S_{11}(\tau)(S_{12}(\tau)+S_{21}(\tau)\Big[(2S_{11}(\tau)S_{22}(\tau) + S_{21}^2(\tau) + S_{12}^2(\tau)) \\
& S_{11}S_{22} + 2\big(S_{11}(\tau)S_{22}(\tau) + S_{12}(\tau)S_{21}(\tau)\big)S_{12}^2\Big].
\end{aligned}
$$

Note that the latter expressions are complicated, but easy to implement as demonstrated in the remaining sections. Again for this case where $p = 2$, the approximation $\frac{N-1}{N+1} = 1 - \frac{2}{N} + o(1/N)$ was used in the expressions of the mean and variance of $\hat{B}_2$.

4.6 PARTICULAR CASE: MULTIDIMENSIONAL EMBEDDING OF A SCALAR PROCESS

In this section, we consider the particular case where the multivariate process consists of the embedding of a scalar process. More precisely, we assume that

$$
\mathbf{x}(n) = \begin{pmatrix} x_1(n) \\ \dots \\ x_p(n) \end{pmatrix} = \begin{pmatrix} y(n\delta+1) \\ \dots \\ y(n\delta+p) \end{pmatrix}.
$$

where $y(k)$ is a scalar wide-sense stationary process of correlation function $C(\tau) = \mathbb{E}\{y(k)y(k-\tau)\} = S_{11}(\tau/\delta)$. Note that now, because of the particular form of $\mathbf{x}(n)$, we can exploit the translation invariance by remarking that $S_{ab}(\tau) = \mathbb{E}\{x_a(n\delta)x_b(n\delta-\tau\delta)\}$ implies $S_{ab}(\tau) = C(\tau\delta + a - b)$, for $1 \leqslant a, b \leqslant p$.

To keep results as concise as possible, we assume the notation $\gamma_i(\tau) = C(\tau\delta + i)$, and the shortcut $C_j = C(j)$. The main goal targeted by defining these multiple notations is to obtain more compact expressions.

4.6.1 BIVARIATE EMBEDDING

The bivariate case is more difficult but the expressions still have a simple form:

$$\boxed{\mathbb{E}\{\hat{B}_2\} \approx 8 - \frac{16}{N} - \frac{4}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)q_1(\tau)}{(C_0^2 - C_1^2)^2}} \tag{157}$$

$$\boxed{Var\{\hat{B}_2\} \approx \frac{64}{N} + \frac{16}{N^2}\sum_{\tau=1}^{N-1}\frac{(N-\tau)q_2(\tau)}{(C_0^2 - C_1^2)^4}} \tag{158}$$

with $q_1(\tau)$ and $q_2(\tau)$ defined below, where γ_i stands for $\gamma_i(\tau)$:

$$\begin{aligned} q_1(\tau) = &\left[(\gamma_1 + \gamma_{-1})^2 + 8\gamma_0^2\right] C_0^2 - 12 C_0 C_1 \gamma_0(\gamma_1 + \gamma_{-1}) \\ &+ \left[2(\gamma_1 + \gamma_{-1})^2 + 4\gamma_0^2\right] C_1^2, \end{aligned} \tag{159}$$

$$\begin{aligned} q_2(\tau) = &\left[8(\gamma_0^2 - \gamma_1\gamma_{-1})^2 + 3(\gamma_1^2 - \gamma_{-1}^2)^2 + 12\gamma_0^2(\gamma_1 + \gamma_{-1})^2\right] C_0^4 \\ &+ 4\left[8\gamma_0^4 + 3\,(5\gamma_0^2 + \gamma_1\gamma_{-1})(\gamma_1 + \gamma_{-1})^2 - 4\gamma_1\gamma_{-1}(\gamma_0^2 + \gamma_1\gamma_{-1})\right] C_0^2 C_1^2 + \\ &8\left[\gamma_0^4 + 4\gamma_0^2\gamma_1\gamma_{-1} + \gamma_1^2\gamma_{-1}^2\right] C_1^4 - 24 C_0 C_1 \gamma_0(\gamma_1 + \gamma_{-1})\left[(2\gamma_0^2 + \gamma_1^2 + \gamma_{-1}^2) C_0^2\right. \\ &\left. + 2(\gamma_0^2 + \gamma_1\gamma_{-1}) C_1^2\right]. \end{aligned} \tag{160}$$

The results we obtain match the expressions derived in [31]. We have also noticed an error in the expression of the variance reported in the first line of the formula (16) of [31]. The exact computation for the trivariate embedding case have also been conducted; but because of their lengthy expressions (especially that of the variance), they are not detailed here. They can be found in [45].

4.7 THE TEST IN PRACTICE

This section is devoted to the practical implementation of the normality test. Given a dataset from a process $\mathbf{X}$, we detail how the test is applied from start to finish.

Remark 4.7.1 *If* $\mathbf{X}$ *has a dimension* $d > 2$*, apply the methodology detailed in sub-section 4.8.2. Henceforth, we suppose that* $d \leqslant 2$*.*

- Given the realizations $\{\mathbf{x}(1), \ldots, \mathbf{x}(N)\}$ of $\mathbf{X}$, estimate as follows the quantities:

$$\begin{aligned}
\hat{B}_d &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}(n)^T\hat{\mathbf{S}}^{-1}\mathbf{x}(n)\right)^2 \\
\hat{\mathbf{S}} &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}(k)\mathbf{x}(k)^T \\
\hat{S}_{ab}(\tau) &= \frac{1}{N}\sum_{k=1}^{N-\tau}x_a(k)x_b(k+\tau), \quad a, b \in \{1, \ldots, d\}
\end{aligned}$$

Then, let us apply:

1. First, we choose the nominal level of the test , that is [139]:

$$\alpha = \mathbb{P}(\text{choose } \bar{\mathcal{H}_0}|\mathcal{H}_0 \text{ is true}) \tag{161}$$

2. Define the ratio $z = \hat{B}_{(.)} - \mathbb{E}\{\hat{B}_{(.)}\}/\sqrt{\text{Var}\{\hat{B}_{(.)}\}}$ for $\mathbb{E}\{\hat{B}_{(.)}\}$ either equal to Equation (153) for a scalar colored process or (155) for a bivariate colored process. Similarly, $\text{Var}\{\hat{B}_{(.)}\}$ is either equal to (154) or (156) according to d. To compare with Mardia's Test, henceforth denoted $\hat{B}_{1,i.i.d}$, we use the results of Theorem 2.

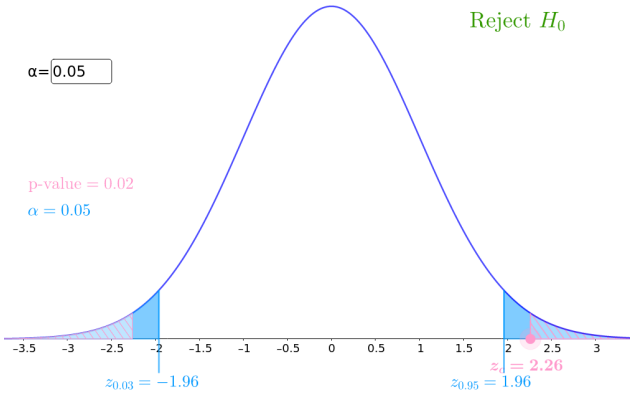
3. Knowing the limiting distribution of z gives us access to the p-values computed as:

$$p_z = 2(1 - F(z)) \tag{162}$$

   Where Φ denotes the cumulative distribution function (cdf) of a standard normal.

4. We reject $\mathcal{H}_0$ at a significance level α if $p_z < \alpha$.

Remark 4.7.2 *Note that Equations (153), (155), (154) and (156) involve the exact covariance functions. However, the covariance function* $S_{ab}(\tau)$ *is replaced in practice by its sample counterpart.*



(a) Rejection rejection in blue of the normality test $\alpha = 5\%$

4.8 PERFORMANCE COMPARISON ON COLORED COPULA

4.8.1 COLORED COPULA

The aim of this section is to generate a *colored* and *bivariate* non-Gaussian process, whose two marginals are Gaussian. For this aim, we briefly introduce a classical framework called Copula, which is simple to implement for defining multivariate distributions with controlled joint distribution function.

For their ease to generate in dimension $d \geqslant 2$, we choose in this work to use Archimedean copula:

Definition 4.8.1 *A d-dimensional copula $\mathcal{C}_\theta$ is called Archimedean if it allows the representation [110]:*

$$\mathcal{C}_\theta(\mathbf{u}) = \psi_\theta(\psi_\theta^{-1}(u_1) + \psi_\theta^{-1}(u_2) + \ldots + \psi_\theta^{-1}(u_d)), \mathbf{u} \in [0,1]^d \tag{163}$$

where $\psi_\theta : [0,\infty) \to (0,1]$ is called an Archimedean generator, and the parameter θ controls the *spatial* dependence between variables.

Theorem 8 *(Sklar's theorem 1959)*

$$F_{X_1,X_2}(x_1,x_2) = \Pr(X_1 \leqslant x_1, X_2 \leqslant x_2) = \mathcal{C}(F(x_1), G(x_2)) \tag{164}$$

where F_{X_1,X_2} is the joint cumulative distribution function (cdf) of (X_1,X_2), and F (resp. G) is the cdf of X_1 (resp. X_2). If F, G are continuous, then $\mathcal{C}$ is unique, and is defined by:

$$\mathcal{C}(u_1,u_2) = F_{X_1,X_2}(F^{-1}(u_1), G^{-1}(u_2)). \tag{165}$$

This theorem guarantees that there is a *unique* copula – called the Gaussian copula $\mathcal{C}_R$ – that produces the bivariate Gaussian distribution, fully specified by the correlation matrix $\mathbf{R}$:

$$\mathcal{C}_R(u,v) = \int_{-\infty}^{F^{-1}(u)} \int_{-\infty}^{F^{-1}(v)} \frac{1}{2\pi(1-R_{12}^2)^{1/2}} \exp\left\{\frac{s^2 - 2R_{12}st + t^2}{2(1-R_{12}^2)}\right\} ds\, dt$$

where F^{-1} is the inverse of the cumulative distribution function of the standard normal distribution. Hence, non Gaussian distributions with Gaussian marginals can easily be sampled by using other types of copulas. Namely here, Clayton and Gumbel bivariate copulas are used as examples:

$$\begin{aligned} \textit{Clayton: } \mathcal{C}_\theta(u,v) &= \max\{u^{-\theta} + v^{-\theta} - 1; 0\}, \theta \in [-1,\infty)\backslash\{0\} \\ \textit{Gumbel: } \mathcal{C}_\theta(u,v) &= \exp\Big\{-(-\log(u)^\theta + -\log(v)^\theta)^{\frac{1}{\theta}}\Big\}, \theta \in [1,\infty) \end{aligned}$$

Moreover we are interested in colored processes, since Sklar theorem does not impose *independence* of any variable u or v of $\mathcal{C}_\theta(u,v)$, we introduce *time-dependency* between samples by applying an auto-regressive filter on each marginal before constructing the copula. Note that the normal distribution is stable by linear transformation, thus the normality of marginals is preserved. This leads to the following algorithm:

SAMPLING AN ARCHIMEDEAN COPULA.

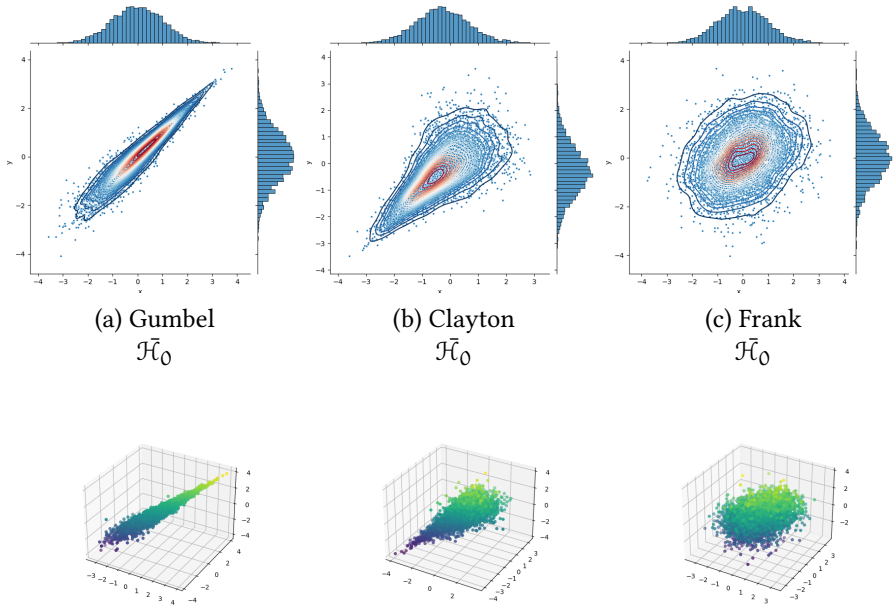
1. Sample i.i.d $\eta_i \sim \mathcal{N}(0,1)$, $i \in \{1,\dots,d\}$
2. Correlate η_i's using a first order auto-regressive filter:

   $$y_i(n) = 0.8y_i(n-1) + \eta_i(n)$$

   Note that the first $n_{drop} = 1000$ samples are dropped to alleviate start-up effects ($y_i(0) = \eta_i(0)$).
3. Transform $u_i = F(y_i)$ for $i \in \{1,\dots,d\}$, where F denotes the cumulative distribution of the Gaussian distribution. Note that u_i's are uniform on $[0,1]$.
4. Sample $V \sim \mathcal{LS}^{-1}(\psi_\theta)$ where $\mathcal{LS}^{-1}$ denotes the inverse Laplace-Stieltjes transform of ψ_θ.
5. Return $(u'_1, u'_2, \dots, u'_d)$, where $u'_i = \psi(-\log(u_i)/V)$
6. Transform u'_i to obtain Gaussian standard marginals as the following:

   $$x_i(n) = F^{-1}(u'_i(n)) \tag{166}$$

The above algorithm is a slight modification to the one due to Mashall, Olkin (1988)[70]. The produced d marginals are now colored and Gaussian, but their joint distribution is not. In the remainder of this paper, we precisely use Gumbel ($\theta = 5$) and Clayton ($\theta = 2$) copula.



(a) Gumbel $\mathcal{H}_0$ (b) Clayton $\mathcal{H}_0$ (c) Frank $\mathcal{H}_0$

Figure 8: Examples of Archimedean Copula with Gaussian Marginals

SIMULATION STUDY

For a given copula $\mathcal{C}$, we perform $M = 2000$ realizations of $\mathbf{x}(n) = (x_1(n), x_2(n))^T$ of total length $N = 1000$. We seek to compare the performance of the normality test when applied on a one-dimensional marginal (we choose arbitrarily $\mathbf{x}_1$) with the joint normality test applied on the bivariate variable $\mathbf{x}$. First, the p-values of the two-sided tests are computed using the limiting distribution of the ratio:

$$z = \frac{\hat{B}_{(.)} - \mathbb{E}\{\hat{B}_{(.)}\}}{\sqrt{\mathrm{Var}\{\hat{B}_{(.)}\}}}$$

Recall that this statistic is standard normal for large N. This gives us access to the the p-values computed as $p = 2(1 - F(z))$. The significance rate (or level of the test) is fixed at $\alpha = 5\%$ or $\alpha = 10\%$. For any p_z smaller than α, it is considered heuristically that the test rejected $\mathcal{H}_0$. The empirical rejection rates, defined by $\frac{\text{Number of rejections}}{M}$ for each statistic $\hat{B}_{1,\text{i.i.d}}$, $\hat{B}_1$ and $\hat{B}_2$ are reported in Table 2.

Test	Gaussian $R_{12} = 0.8$		Clayton $\theta = 2$		Gumbel $\theta = 5$	
statistic	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$
$\hat{B}_{1,i.i.d}$	0.1660	0.2460	0.1011	0.1651	0.1189	0.1930
$\hat{B}_1$	0.0450	0.0730	0.1060	0.1701	0.0390	0.0860
$\hat{B}_2$	0.0480	0.0801	0.9890	0.9920	0.9920	0.9960

Table 2: Empirical Rejection rate at two significance levels : $\alpha = 5\%, 10\%$

- Mardia's test $\hat{B}_{1,i.i.d}$: The one-dimensional marginal being Gaussian (under $\mathcal{H}_0$), the empirical rejection rate is expected to be around the pre-specified nominal level (either 5% or 10%). The rejection rate surpasses the nominal level. That $\hat{B}_{1,i.i.d}$ over-rejects $\mathcal{H}_0$ is due to the one-dimensional marginal being time-correlated (not independently distributed). Such observation was already formulated by [104] and [54] who showed that the correlation among samples is confounded with lack of Normality.
- $\hat{B}_1$ has an empirical rejection rate around the nominal level because the expressions involved take into account the serial dependence between samples $x_1(1), \ldots, x_1(N)$.
- $\hat{B}_{1,i.i.d}$ and $\hat{B}_1$ can only test a one-dimensional marginal, which is Gaussian (under $\mathcal{H}_0$) therefore they are always conservative and mis-detect the non-Gaussianity of the bivariate process $\mathbf{x}$.
- $\hat{B}_2$: The rejection rates do not differ substantially from the nominal level when data is distributed according to bivariate Gaussian. For the Gumbel and Clayton copulas (under $\bar{\mathcal{H}_0}$), this test has very high rejection rates, which confirms the necessity of taking into account the full dimension to design a powerful test.

4.8.2 THE MULTI-VARIATE CASE

We propose the following methodology to deal with the general d-variate case.

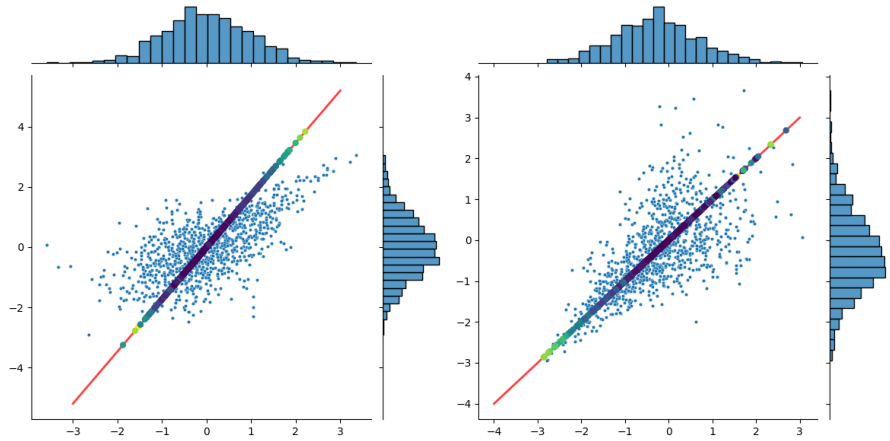
In the same spirit as [111], using the property that Gaussianity is stable by linear transformation, we can randomly project the initial d-variate observations on a bivariate subspace (plane), and test the joint normality of this two-dimensional representation.

LOW-DIMENSIONAL PROJECTION. We study the performance of the proposed test statistic on a low-dimensional (either 1 or 2) projection

of the initial p-variate data. For a given copula $\mathcal{C}_\theta$, we carried out the following simulations:

- Given one set of bivariate observations $(x_1(n), x_2(n))$ of total length $N = 1000$, they are projected $M = 5000$ times onto the arbitrary vector $\mathbf{u}$ with coordinates $(\sin(\varphi), \cos(\varphi))$. φ is sampled from a uniform distribution on $[0, \pi]$ denoted $\mathcal{U}(0, \pi)$. **Fig.**9 shows an illustrative example with two copulas.

- Given one set of trivariate observations of total length $N = 1000$, the points are projected arbitrarily $M = 5000$ times onto the plane defined by two angles $\theta \sim \mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$ (the angle between the z axis and the new plane) and $\varphi \sim \mathcal{U}(0, \pi)$ (measured between the x axis and the vector $\mathbf{u}$ inside the plane). **Fig.** 10 gives two illustrative examples of this procedure.



(a) Gumbel Copula (b) Clayton Copula

Figure 9: Two examples of projecting **bivariate** realizations (in blue) onto the direction in red defined by the angle φ. Realizations are generated using Clayton copula (on the left), and Gumbel copula (on the right). The distribution of canonical marginals are both Gaussian as illustrated by histograms but the bivariate distribution is clearly not Gaussian.

SCALAR PROJECTION

- $\hat{B}_1$ and $\hat{B}_{1,i.i.d}$ perform very poorly when used on arbitrary one-dimensional projections of the Gumbel copula. The test power does not surpass 25%.

- For the Clayton copula, whose tails are asymmetric, the test has a better power than the Gumbel copula. Although this ob-

Time correlated	$\hat{B}_1$		$\hat{B}_{1,i.i.d}$	
	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$
Gumbel	0.1250	0.1328	0.1242	0.1316
Clayton	0.661	0.72	0.652	0.713

Table 3: Empirical rejection rates of the test applied to a one-dimensional projection of a bivariate process with time-correlated Gaussian marginals.

Time independent	$\hat{B}_1$		$\hat{B}_{1,i.i.d}$	
	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$
Gumbel	0.2134	0.2510	0.2082	0.2406
Clayton	0.717	0.76	0.701	0.752

Table 4: Empirical rejection rates of the test applied to a one-dimensional projection of a bivariate process with independent (in time) standard normal marginals.

servation is less demonstrative, we keep those results to further compare them with the bivariate test statistic.

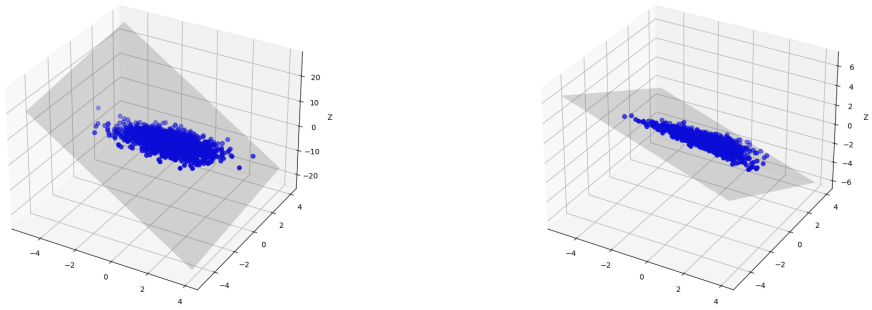
- Since we only use first-order auto-regressive filters, there is no substantial difference in the performance of $\hat{B}_1$ compared to $\hat{B}_{1,i.i.d}$; this comparison is not of interest to us, because the bias induced by using tests assuming independence on colored processes has already been observed and studied in the literature [54] and [104].
  However, it is interesting to compare Tables 3 and 4; we see that the overall performance of the test statistics tends to decrease when marginals are **time-correlated**.

Arbitrary 2-D projections	$\hat{B}_2$	
	$\alpha = 5\%$	$\alpha = 10\%$
Gumbel	0.9516	0.9574
Clayton	0.9701	0.9882

Table 5: Empirical rejection rates of the test applied jointly to arbitrary 2-D projections.

BIVARIATE PROJECTION

- **Table 5** shows performances obtained with $\hat{B}_2$ when applied to colored processes. Contrary to $\hat{B}_1$, performances do not de-



(a) Gumbel Copula (b) Clayton Copula

Figure 10: Two examples of projecting **trivariate** realizations onto the plane (in light grey). Realizations are generated using Clayton copula (on the left), and Gumbel copula (on the right). The distribution of the two-dimensional projection is clearly not Gaussian.

Time-correlated marginals	$\hat{B}_2$	
	$\alpha = 5\%$	$\alpha = 10\%$
Gumbel	0.9492	0.9556
Clayton	0.8540	0.87

Table 6: Empirical rejection rates of the test applied to the two-dimensional projection of a trivariate process with time-correlated marginals.

crease with time-correlation. Furthermore, the power of the 2-D test based on $\hat{B}_2$ is not affected by a rotation in the plane (implemented by two scalar projections onto two orthogonal axes). This is illustrated by **Table 5**, which reports the results averaged over 5000 random rotations.

- One would expect the same problem of misdetections to occur when projecting trivariate observations sampled from Gumbel copula. Yet, in **Table 6** we show that the joint normality, even on a low representation of the data, is able to detect the non-Gaussianity of the process.

4.9 CONTRIBUTIONS

We derived the limiting distribution of the bivariate kurtosis under $\mathcal{H}_0$. This kurtosis test is intended to test the *joint normality* when multivariate time-series are observed. The complexity of the calcula-

tion increases with d very fast, therefore exact expressions were only conducted and reported up to $d \leqslant 2$. However, we propose and implement a strategy based on random projections. For trivariate processes, we have confronted the performances of a strategy based on 1D vs. 2D projections and concluded that the latter is more robust to misdetections. We will carry on this exercise of testing the performances of the proposed results as a practical *online detector*. In Chapter 5, we propose an operational detector and evaluate it on synthetic data. Future studies were then conducted in 6 for high-dimensional real data.

5

THE NORMALITY TEST AS A SEQUENTIAL DETECTOR

ABSTRACT

We anticipate that real-data will be very complex in nature, in the sense that it will be non-stationary, and composed of both linear and non-linear components. The workaround to handle non-stationarity is to use the idea of local stationarity by means of an exponential averaging mechanism. As for the second point, for now, we conduct numerical experiments on data generated by a linear filter (the multi-dimensional VAR). The aim is to put to test the performances of the normality test as sequential detector with and without a first stage of recursive pre-whitening. Its robustness to mis-specifications in the linear model is studied. We also propose an extension for the general multivariate case by means of bivariate random projections.

Contents

5.1 SEQUENTIAL CHANGE DETECTION

Detecting changes in the distribution of a stochastic process is a long-standing problem and a myriad of methods has been proposed; see e.g [10]. Our concern is the detection of non-Gaussian signals in a *Gaussian* background, *i.e. Normality tests*. The framework considered here is one in which time-series are recorded on d sensors (typically $d = 2$ or 3 in many applications). Moreover, we are interested in the *online* problem of reacting to a change as quickly as possible after it occurs, also known as *sequential detection* problem [12, 117]. A popular sequential detector is based on the Likelihood Ratio and termed the CUSUM test [69, 115]. In this work, we focus on translating the results of Chapter 4 into an operational detector. As we have seen in Chapter 3, The case of i.i.d (scalar or d-variate) processes has received significant attention. On the other hand, few tests concern the case of dealing with time-series from multiple sensors that in practical applications cannot be considered to be i.i.d, a case we will refer to as *n.i.d* or *colored*.

Additionally, we are interested in testing the normality of *unobserved* regression residuals. More precisely, define the Multi-dimensional Auto-Regressive [93] (or Vector AR of order p denoted VAR(p)) model to describe the statistical behavior of the $d \times 1$ vector of observation $\mathbf{x}(i)$ for $i = 1, \dots, N$:

$$\mathbf{x}(i) = \sum_{k=1}^{p} \boldsymbol{\beta}_k \mathbf{x}(i-k) + \boldsymbol{\epsilon}(i) \tag{167}$$

where $\boldsymbol{\beta}_k$ is a $d \times d$ matrix of unknown parameters and $\boldsymbol{\epsilon}(i)$ is the ith unobservable residual assumed zero-mean and i.i.d. An additional assumption is that residuals are drawn from a normal distribution. The drawbacks of violating the latter assumption has been studied, for instance [71] showed that the ordinary least squares method, which is usually used to estimate $\{\boldsymbol{\beta}_k\}_{1 \leqslant k \leqslant p}$, is sub-optimal for heavy-tailed distributions. Thus, it is important to validate this assumption of normality in this linear model.

Moreover, since the residuals are estimated in practice, we expect that errors in the model specification and estimation will impact the whitening performance of the filter and the estimated innovation process $\widehat{\boldsymbol{\epsilon}}$ could no longer be considered i.i.d. In this case, the normality tests designed for i.i.d processes become biased as shown in [104], highlighting the importance of deriving our joint normality test for variables that are not statistically independent.

Within this framework, we concentrate on the *Multivariate Kurtosis* (MK) defined in (140). In Chapter 4, we calculated the power of

this test variable in the colored case, we also studied its performances when the observed multivariate process was projected onto an arbitrary subspace of low dimension (typically 1-D or 2-D), in particular for time series generated by colored copulas.

The main contributions of this chapter are the following:

- We compare the performances of the normality test with and without linear prewhitening, *i.e.* using $\mathbf{x}(n)$ or $\boldsymbol{\epsilon}(n)$.
- We observe the impact of projecting 3-D observations and their *regression residuals* onto an arbitrary plane (2-D projection) or direction (1-D projection).
- For sake of time and memory effectiveness, the test statistics and the regression function are estimated recursively, by using both exponential averaging and the Recursive Least Squares (RLS) algorithm.

5.2 ALGORITHMIC IMPLEMENTATION

The proposed test statistic is computationally efficient i.e it can be easily computed over a sliding window or by using the exponential weighting technique to test the normality of the estimated regression residuals available at time t and assumed to follow a Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma}(t))$. The residuals are estimated using the recursive least squares estimation method detailed in Equation 2.2.1. Let $0 < \lambda_2 < 1$:

$$\boldsymbol{\Sigma}(t) = \lambda_1 \boldsymbol{\Sigma}(t-1) + (1-\lambda_1)\hat{\boldsymbol{\epsilon}}(t)\hat{\boldsymbol{\epsilon}}(t)^T \tag{168}$$

$$\hat{B}_d(t) = \lambda_2 \hat{B}_d(t-1) + (1-\lambda_2)(\hat{\boldsymbol{\epsilon}}(t)^T \widehat{\boldsymbol{\Sigma}}^{-1}(t)\hat{\boldsymbol{\epsilon}}(t))^2. \tag{169}$$

A WORD ON THE CHOICE OF λ_1, λ_2 Forgetting factors λ_1 and λ_2 (for 2nd and 4th order statistics respectively) are usually chosen by a rule of thumb, depending on the time-scale of the change. To give a better intuition of this factor, one can calculate the length N of a uniform sliding window that would yield the same estimator's variance. Let $a, b \in \mathbb{R}$, for $\boldsymbol{\epsilon} \underset{\text{i.i.d}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$, the following estimator is unbiased when $a + b = 1$.

$$\boldsymbol{\Sigma}(t) = b\boldsymbol{\Sigma}(t-1) + a\hat{\boldsymbol{\epsilon}}(t)\hat{\boldsymbol{\epsilon}}(t)^T$$

$$\mathbb{E}\{\boldsymbol{\Sigma}(t)\} = a\sum_{k=0}^{\infty} b^k \mathbb{E}\{\boldsymbol{\epsilon}(t-k)\boldsymbol{\epsilon}(t-k)^T\} \quad (170)$$

$$\mathbb{E}\{\boldsymbol{\Sigma}(t)\} = \frac{a}{1-b}\boldsymbol{\Sigma} \quad (171)$$

$$\mathrm{Cov}(\Sigma_{ab}, \Sigma_{cd}) = \frac{2a}{2-a}\Sigma_{ab}\Sigma_{cd} \quad (172)$$

Additionally, for a Gaussian process, the variance of this estimator on N samples is of order N^{-1}:

$$\mathrm{Cov}(\Sigma_{ab}, \Sigma_{cd}) = \frac{1}{N^2}\sum_{m,n=1}^{N} [2]\mu_{ac}^{nm}\mu_{bd}^{nm} \quad (173)$$

We recall that $\mu_{ab}^{ij} = \mathbb{E}\{\epsilon_a(i)\epsilon_b(j)\}$, For Gaussian i.i.d $\mathrm{Cov}(\Sigma_{ab}, \Sigma_{cd}) \propto \frac{2}{N}$, to get a rough estimate about the number of samples included in the estimate of the exponential averaging, we have: $N \propto \frac{2}{a} = \frac{2}{1-\lambda_1}$. Concerning the choice of λ_2 for the recursive estimation of $\hat{B}_2$, we show by means of numerical simulation, that we also obtain: $N \propto \frac{2}{a} = \frac{2}{1-\lambda_2}$.

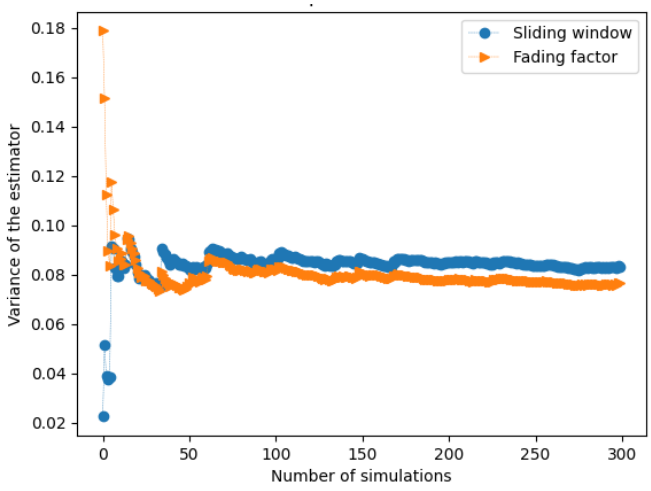


Figure 11: 300 Monte-Carlo simulations to verify that the empirical variance of the sliding window estimator and exponential averaging estimator of $\hat{B}_2$ are equal for $N \propto \frac{2}{1-\lambda_2}$

The algorithm for sequentially detecting changes reads:

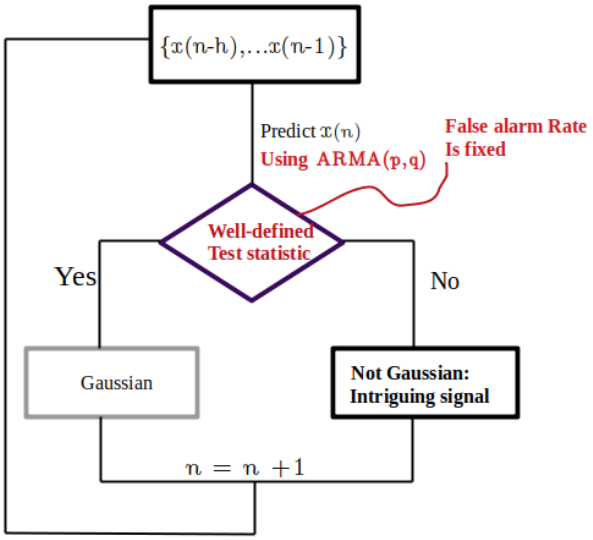


Figure 12: Illustration of the detection workflow

Algorithm 1 Sequential Change detector

Require: $p \geqslant 1, 0 < \lambda_1, \lambda_2 < 1, \alpha$
  Initialization: $\Gamma(p) \leftarrow I_{dp}$, $\hat{B}_d(p) \leftarrow 0$, $\Sigma(p) \leftarrow I_d$
  for $p+1 \leqslant t \leqslant N$ **do**
    Update $\Gamma^{-1}(t)$ ▷ using (79)
    Update $\{\widehat{\beta}_k(t)\}_{1 \leqslant k \leqslant p}$ ▷ using (81)
    Compute $\hat{\epsilon}(t) = x(t) - \sum_{k=1}^{p} \widehat{\beta}_k x(t-k)$
    Compute $z = (\hat{B}_d(t) - \mathbb{E}\{\hat{B}_d(t)\})/\sqrt{\mathbb{V}ar(\hat{B}_d(t))}$
    if $2(1 - F(z)) < \alpha$ **then**
      Change is detected
    else
      No change
    end if
  end for

5.3 COMPUTER RESULTS

A set of of Monte Carlo simulations is presented to compare the power of the proposed normality test when applied directly on data or on regression residuals. Then, we study the performance of the test statistic on a low-dimensional (2-D or 1-D) projection of the initial multivariate data. As a final illustration of the effectiveness of our method, we apply the change detection Algorithm presented in subsection 5.2 on

synthetic colored data undergoing an abrupt change in its distribution.

5.3.1 APPLYING THE TEST ON COLORED DATA

5.3.1.1 *Directly on data*

$M = 2000$ simulations are considered, each being based on a sequence of $N = 1000$ samples. First, we simulate 1-D AR(p) processes (for $p \in \{4, 14, 20\}$). The AR coefficients are computed such that the equivalent filter is low-pass, with band pass equal to .25 (normalized freq).

For each simulation a 2-D Gaussian (or Uniform) AR(p) process is constructed by time embedding : $\mathbf{x}(t) = \{x(2t), x(2t+1)\}^T$. Then, both Mardia's test derived for i.i.d. samples (denoted $\hat{B}_{1,i.i.d}$) and the test for colored samples, whose statistics are defined by equations (153, 154) for colored samples (denoted $\hat{B}_1$) are applied on the marginals of the 2-D process, and compared. Finally, the statistics derived for n.i.d. bi-variate data $\hat{B}_2$, described by equations (155, 156) is also applied to the 2-D process. The obtained empirical rejection rates of hypothesis H_0 are computed as $\frac{\#\text{Rejections}}{M}$, for a test level $\alpha = 5\%$. The results are summarized in Tables 7 and 8.

	AR(4)		AR(14)	
Test Statistic	Gaussian	Uniform	Gaussian	Uniform
$\hat{B}_{1,iid}$	0.067	1.	0.123	0.512
$\hat{B}_1$	0.052	0.99	0.045	0.456
$\hat{B}_2$	0.06	1.	0.065	0.88

Table 7: Empirical Rejection Rates for 2000 simulations for $\alpha = 5\%$ significance level with the $\hat{B}_{1,i.i.d}, \hat{B}_1, \hat{B}_2$ test applied directly on AR(p) data with $p = 4, 14$

- It seems that $\hat{B}_{1,i.i.d}$ has a better detection power than $\hat{B}_1$. As a matter of fact, by comparing the formulas of their variance in (Theorem 2) and 154), we can see that that the variance of $\hat{B}_{1,i.i.d}$ is underestimated for colored processes, consequently, the latter over-rejects the hypothesis of Gaussianity. This is noticeable even for Gaussian AR(p) process (Under H_0), with rejection rates that surpass the nominal level 5%.

- Both scalar tests $\hat{B}_{1,\text{i.i.d}}$ and $\hat{B}_1$ perform poorly compared to the joint normality test statistic $\hat{B}_2$.
- When the correlation tails last longer (Table 8), the overall performance of the test statistics tends to decrease.

	AR(20)	
Test Statistic	Gaussian	Uniform
$\hat{B}_{1,\text{iid}}$	0.228	0.430
$\hat{B}_1$	0.047	0.399
$\hat{B}_2$	0.06	0.688

Table 8: Empirical Rejection Rates for 2000 simulations for $\alpha = 5\%$ significance level with the $\hat{B}_{1,\text{i.i.d}}, \hat{B}_1, \hat{B}_2$ test applied directly on AR(20) process

5.3.2 ON REGRESSION RESIDUALS

We now study the power of the normality tests on *estimated* regression residuals. We utilize the Least Squares method to obtain $\hat{\epsilon}$. The simulation procedure and the tests are the same as those described in the previous paragraph 5.3.1.1. We study the case where the order p of the generated AR(p) process is known (we choose $p = 20$), and the case where the order is misspecified ($\hat{p} = 9$). The empirical rejection rates are summarized in Table 9.

	On Residuals of AR(20) $p = 20$		On Residuals of AR(20) $\hat{p} = 9$	
Test Statistic	Gaussian	Uniform	Gaussian	Uniform
$\hat{B}_{1,\text{iid}}$	0.06	1.	0.064	0.582
$\hat{B}_1$	0.05	1.	0.051	0.429
$\hat{B}_2$	0.055	1.	0.06	0.850

Table 9: Empirical Rejection Rates for 2000 simulations for $\alpha = 5\%$ significance level with the test statistics $\hat{B}_{1,\text{i.i.d}}, \hat{B}_1, \hat{B}_2$ applied on estimated regression residuals using OLS method, with known and misspecified order

	3-D VAR(5)	
2-D projection	Gaussian	Uniform
$\hat{B}_2$	0.051	0.986

	3-D VAR(5)	
1-D projection	Gaussian	Uniform
$\hat{B}_1$	0.056	0.529

Table 10: Empirical rejection rates of the test statistics applied to a low representation of 3-D VAR(p) process

- If the model is perfectly known, then all test statistics perform well on the well estimated regression residuals. However, if the model order is under estimated, some important time correlation remain, and scalar tests perform poorly compared to the joint normality test, as expected.

5.3.3 RANDOM PROJECTIONS

We simulate a 3-D process VAR(p), $p \in \{5, 20\}$, of length $N = 1000$ following equation (167), where the inputs $\epsilon(t)$ are i.i.d with distributions either multivariate standard normal $\mathcal{N}(0, 1)$ or multivariate $\mathcal{U}(-2, 2)$.

Then $M = 2000$ different projections on an arbitrary plane (2-D projection) going through the origin of the 3-D space are computed, corresponding to as many 2-D time series. For comparison, the same set of observations is also projected M times on an arbitrary direction (1-D projection). Eventually, we run the same set of experiments on estimated regression residuals estimated by ordinary least squares method (OLS). The results are reported below.

	3-D VAR(20)	On residuals $\hat{p} = 10$
1-D proj., $\hat{B}_1$	0.250	0.41
2-D proj., $\hat{B}_2$	0.580	0.9

Table 11: Empirical rejection rates of the test statistics applied to a low representation of 3-D VAR(20) process with uniform inputs and its estimated regression residuals with a VAR(10) model

- The test $\hat{B}_1$ applied directly on 1-D projections of either the observations or its residuals computed by OLS, performs poorly. This is in accordance with our observations from the preceding experiments.

- Even with a misspecified order, the joint test statistic performs best (empirical power of 90%) on the 2-D low representation of regression residuals, as it is able to account for both temporal and spatial (between coordinates) dependences.

Consider the case where a process ϵ is constituted of i.i.d samples following a standard normal distribution. The process undergoes a change at $n_c = 5000$ in its distribution: samples in the interval $[5000, 10000]$ are now following a uniform distribution $\mathcal{U}(-\sqrt{3}, \sqrt{3})$. The change ends at $n = 10000$ and the samples are again normally distributed. The process ϵ is then (low pass) filtered using an AR(5) model and is now denoted x.

One realization of this process is given in Fig. 13. It is clear from this figure that the change in the distribution is unbeknownst to the human eye.

The change detection algorithm presented in subsection 5.2 is applied to this realization by setting $p = 5, \lambda_1 = 0.99, \lambda_2 = 0.998$, $\alpha = 5\%$ and $\delta = 1$. A 2-D process is obtained by taking $\mathbf{x}(t) = \{x(2t), x(2t+1)\}$.

For comparison, the CUSUM algorithm [12] is applied on the regression residuals by using its recursive form. The instantaneous log-likelihood ratio is computed as:

$$L(t) = -\ln(2\sqrt{3}) + \frac{1}{2}\ln(2\pi) + \frac{1}{2}\hat{\epsilon}^2(t) \tag{174}$$

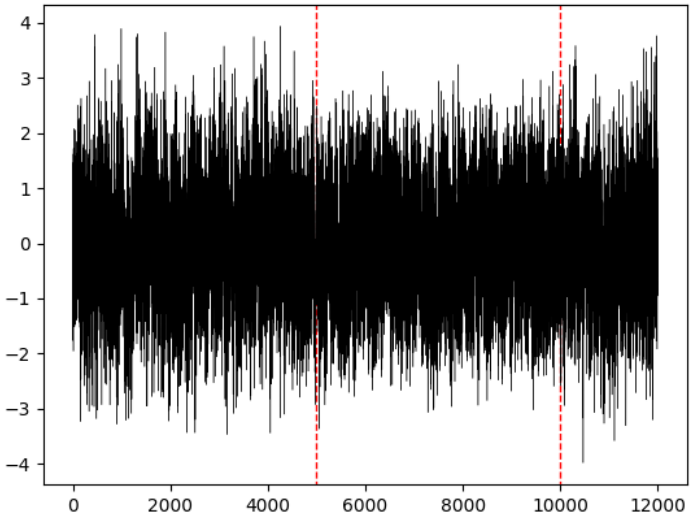


Figure 13: One realization of a Gaussian AR(5) process that undergoes an abrupt change in the distribution of its excitation ϵ (from $\mathcal{N}(0,1)$ to $\mathcal{U}(-\sqrt{3}, \sqrt{3})$). Affected samples are between red dashed lines.

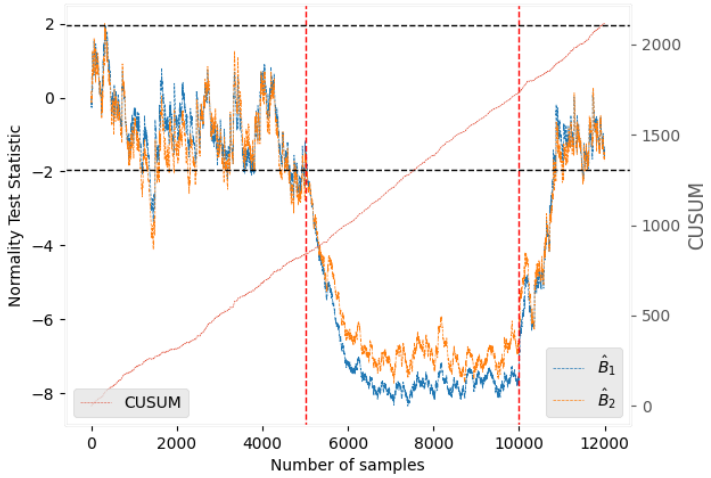


Figure 14: Evolution of the normalized test statistics $\hat{B}_1$ (in blue) and $\hat{B}_2$ (in orange). In red, the evolution of the cumulative sum $s(n) = \sum_{t=1}^{n} L(t)$. Black horizontal dashed lines are the critical values ± 1.96 corresponding to a test power of 95%. Red vertical dashed lines are the beginning and end of an abrupt change in the excitation statistics of an AR(5) process.

- Our proposed test statistics $\hat{B}_{1,2}$ stay in between ± 1.96; in other words they do not reject the null hypothesis of Gaussianity with a false alarm rate of 5%. They grow continuously in absolute value after the change time $n_c = 5000$ and keep rejecting H_0 until the end of the change at $n = 10000$.

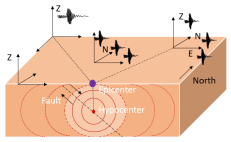
- There is no clear-cut in the cumulative sum algorithm that indicates a change in the distribution of the residuals. In fact, the likelihood ratio in (174) is derived under the hypothesis that the process ϵ is i.i.d. As the latter's values are estimated recursively, they are more likely to have residual correlations between them and the hypothesis of *independence* no longer holds, explaining why the cumulative sum of $L(t)$ keeps increasing.

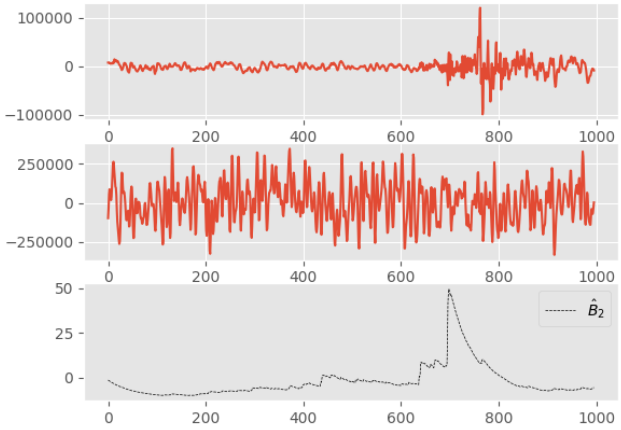
5.4 VALIDATION ON REAL DATA

Finally, our concern is to propose an operational detector in a real environment, we perform two simulations on the response of a single three-axis measuring instrument, and a network of three-axis networks. The real observations are filtered by a vector auto-regression filter of order 15. The residuals are then projected on an arbitrary plane. Finally, we recursively estimate the test statistic and its threshold. Results appear in Figure 15a.

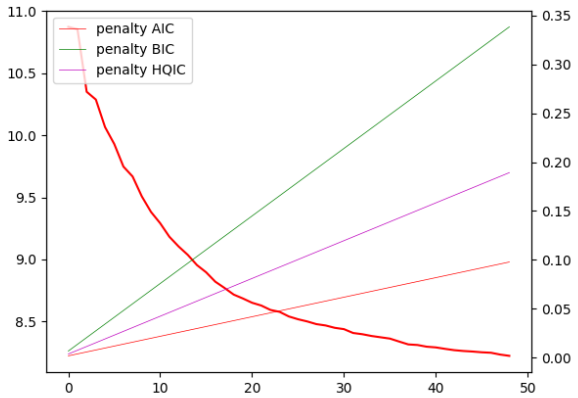
order is estimated using BIC see Figure 15b

Consider now, that we have a network of three-axis sensors:





(a) Top: (horizontal) component of a seismic signal; Middle: observation of the noise added to the seismic signal (SNR= −5dB); Bottom: detection function obtained after two-dimensional projection of the VAR filter residuals



(b) IC w.r.t $1 \leqslant p \leqslant 50$

Figure 15: The arrival of a seismic wave is translated by a peak in the test statistic (a) bottom.

$$x(n,i) \approx \sum_{k=1}^{p} A_{k,i} x(n-k,i) + \epsilon(k,i) \tag{175}$$

where $\{x(n,i) \in \mathbb{R}^d, n = 1,\dots,N\}$ is the set of observations carried out on the sensor or the sub-group of sensors indexed by i. The problem can be summarized as the form of multiple binary tests for all $1 \leqslant i \leqslant N_c$ where N_c is the total number of sensors in the network.

$$\mathcal{H}_0^{(i)} : \epsilon(i) \underset{\text{n.i.d}}{\sim} \mathcal{N}(0, \Sigma(i)) \quad \textit{versus} \quad \bar{\mathcal{H}}_0^{(i)} \tag{176}$$

We randomly project N_p times the residuals of each sensor on a plane ($d = 2$) or on a direction ($d = 1$).

We have $m = N_p \times N_c$ hypotheses $\mathcal{H}_0^{(i)}$, $i = 1, \dots, m$ to test. If all the tests are thresholded with α, the false alarm level is controlled at $m\alpha$ and leads to a large number of false alarms, or *false discoveries*. A first solution consists in thresholding each test by α/m, this correction is due to Bonferroni.

We choose the procedure proposed by Benjamini and Hocheborg (BH) [14] to control the false discovery rate (FDR): the rate of true $\mathcal{H}_0^{(i)}$ wrongly rejected among all rejected hypotheses. Let $0 < \delta \leqslant 1$ be a control parameter for the FDR:

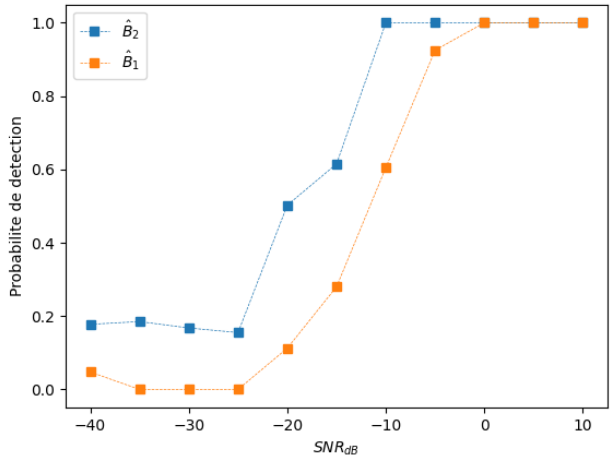
1. Let $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(m)}$ be the ordered p-values.
2. Let $k = \underset{i}{\operatorname{argmin}}\left(p_{(i)} \leqslant \frac{i}{m}\delta\right)$
3. We reject the k null hypothesis $\mathcal{H}_0^{(1)}, \mathcal{H}_0^{(2)}, \dots, \mathcal{H}_0^{(k)}$

The observation is modeled by an additive contamination:

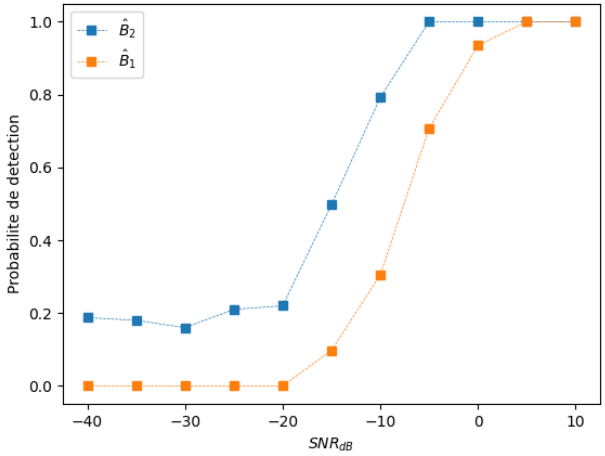
$$\mathbf{x}(n, i) = \text{Diag}\{\mathbf{a}\}\,\mathbf{s}(n, i) + \mathbf{b}(n, i) \tag{177}$$

$\mathbf{s}(n, i) \in \mathbb{R}^3$ est real seismic signal recorded by sensor i from $N_c = 8$ sensors; $\mathbf{a} \in \mathbb{R}^3$ adjusts the Signal to Noise Ratio (SNR); the process $\mathbf{b}$ is:

- *Synthetic colored noise:* zero-mean Gaussian signal filtered by a low-pass VAR(5). This synthetic noise can correspond in real applications to measurement noise.
- We study the impact of the recursive pre-whitening filter's order on the performances of the test for $p = 5$ (correct order, cf. Fig. 16a) and $p = 2$ (mis-specified order cf. Fig. 16b)
- The control parameter of the FDR is $\delta = 5\%$.
- The sample size is $N = 1000$. The number of three axis sensors is $N_c = 8$.
- The residuals obtained by whitening each sensor response are randomly projected $N_p = 5$ times on a plane that passes through the origin of the initial (three-dimensional) space. To compare with the performance of $\hat{B}_1$, the same residuals are projected $N_p = 5$ times onto an arbitrary (random) direction.



(a) Detection power of the univariate and bivariate kurtosis test in the case where the order of the additive Gaussian noise is set to 5 and whitened by a VAR(5)



(b) Detection power of the univariate and bivariate kurtosis test in the case where the order of the additive Gaussian noise is set to 5 and whitened by a mis-specified VAR(2)

SIMULATION RESULTS

- For all simulations, the detector that takes into account the spatial and temporal correlation $\hat{\mathrm{B}}_2$ has a better detection power than its scalar equivalent $\hat{\mathrm{B}}_1$. And this, for very low SNR.
- When the order of the pre-whitening filter is mis-specified, the performance of the $\hat{\mathrm{B}}_1$ based detector degrades. The results of the $\hat{\mathrm{B}}_2$ detector are good for an SNR $\geqslant -15$dB.

5.5 CONTRIBUTIONS

This study demonstrates, on one hand, that testing the joint normality of a two-dimensional projection yields a noticeable increase in the

power of the test to detect departure from joint normality. On the other hand, when data are additionally time-correlated, the overall power of the scalar test tends to decrease. By assuming both spatial and temporal dependence, the bivariate test has better power properties than its univariate counterpart. We have also conducted experiments to assess the impact of the pre-whitening stage on the power of the test. They have evidenced the fact that the test is robust to misspecifications in the model. Hence, we proposed a two-step detection algorithm, in a first stage, data are whitened recursively. Then the MK test is applied in a second stage on available regression residuals in an *online* manner.

Part III

APPLICATIONS IN SEISMOLOGY

The field of seismology is rich with techniques developed for earthquake detection, phase picking, seismic tomography and many other worthy applications in Geodesy. Available seismic data is dramatically increasing both in volume and variety. Hence, reliable Machine learning methods can be a complementary set of tools to extract valuable information from these loads of data. Chapter 6 reviews some applications of Machine Learning in seismology, with a focus on the automated detection problem. We then proceed with reviewing the tools presented in the previous Chapters: the VAR(p), the LSTM model met in Chapter 2, a variant of the scattering transform met in Chapter 1. Finally, and most importantly, the workflow presented in Chapter 5 is put to test on high-dimensional ($d \geqslant 3$) real-data.

6

APPLICATIONS IN SEISMOLOGY

ABSTRACT

ML methods have seen widespread adoption in seismology in recent years. This is due to the fact that seismology is a rich-intensive field with a variety of data i. e. seismograms, GPS, Radar/LiDar images. The spectrum of applications in seismology is also wide, in this work, we focus on two important tasks: detection of unseen seismic waves and phase picking. The detection can be related to both classification or prediction tasks. We review some selected methods that illustrate this. We also tackle the case of detection from multiple stations by using the beamforming technique proposed in [13].

Contents

6.1 APPLICATIONS OF ML IN SEISMOLOGY

One of the first uses of Machine Learning in seismology was to the problem of discrimination of seismic waves from mining explosions and their classification using ANN in [40]. There have been a lot of advances of ML in seismology ever since, see the survey of [85]. We focus on an application that has a long and rich history in geophysics: the detection of seismic tremors. The Gutenberg-Richter law [61] dictates that the cumulative number of earthquakes increases exponentially with decreasing magnitude. These events are severely contaminated by background noise. The background noise can include for example weather effects or large bursts from a nearby human activity. Weak signals are drowned by this ambient noise, detecting them is not an easy task, but accomplishing it will greatly contribute in understanding of the dynamics of Earthquakes. Recent studies are starting to reveal more and more newly detected patterns such as Slow slip events and their associated tremors [123], creep-slip events [116], or Low Frequency Earthquakes[52].

The field of seismology has always been rich with techniques developed for detection and we choose to review the ones relevant to our detection task.

6.1.1 DETECTORS BASED ON STATISTICS

POWER DETECTOR STA/LTA

In seismology, the most commonly used event detection algorithm is the **S**hort-**t**erm **a**verage/**L**ong-**t**erm **a**verage (STA/LTA) detector proposed by [3]: It computes the local Signal to Noise Ratio (SNR) by keeping track of the long-term and short term energy of the signal.

$$\mathrm{STA}(i) = \frac{1}{N_s} \sum_{j=i-N_s}^{i} x(j)^2 \tag{178}$$

$$\mathrm{LTA}(i) = \frac{1}{N_l} \sum_{j=i-N_l}^{i} x(j)^2 \tag{179}$$

Rather than using the complex raw signals, authors in [8] proposed to run the STA/LTA on the envelope function and even proposed the use of an adaptive threshold. This algorithm has the advantage of being rapid and it requires no learning, but it requires the choice of N_s, N_l and a threshold to compare the ratio. The performance of the method

degrades in situations where the SNR is low, and works best if the bandwidth of the desired signal is known.

AUTOMATED DETECTION BASED ON AR

One of the earliest applications of multi-dimensional auto-regressive modeling, met in Chapter 2 as VAR(p), to seismograms dates to 1993; The proposed approaches in [89, 136] were to fit stationary multiple VARs on different segments of the data, and AIC is used to compare the models and detect the onsets of a seismic signal. As discussed in [9], this methodology was not applied to many datasets because the computational complexity is expensive on large data.

DETECTORS BASED ON HIGHER-ORDER STATISTICS

Authors in [125] were first to introduce higher order statistics, the skewness and kurtosis to seismic traces. Subsequently, [9] proposed an automatic phase-picking characteristic function based on the kurtosis, followed by a polarization analysis to detect both the onset of P and S waves. Numerical experiments conducted in [9] evidenced

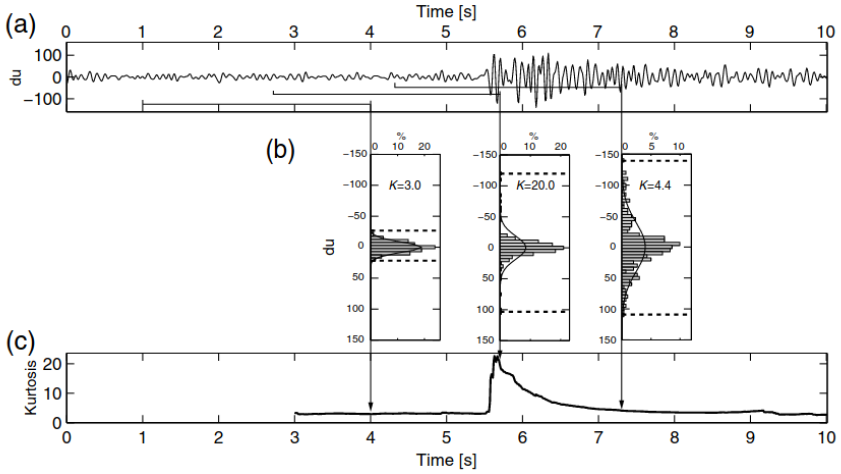


Figure 16: Illustration of the kurtosis based detection on a seismic trace.Image taken from [9].

that the number of picked events and accuracy of the picking are significantly higher than using the STA/LTA detector. The approach we will adopt is very similar in spirit, the main novelties are the use of the bivariate joint normality test, and the results we obtained on its limiting distribution, thus allowing us to control the threshold using a pre-specified false alarm rate.

6.1.2 PROBING EVENTS WITH DEEP LEARNING

There have been a lot of advances in a subset of Machine Learning: Deep learning in seismology, the reader is referred to [85] for examples consolidating this statement. The fact that seismology is a data-rich field has attracted the use of deep learning methods. With no intention of being exhaustive, we review in the following some selected methods that show promising results in automated detection and/or phase picking.

ZOOM ON SCATTERING NETWORK

The following architecture is proposed in [129] for the task of unsupervised classification of seismic signals. At each layer $^{(i)}$, the mother wavelet is used to derive a number of $J^{(i)}Q^{(i)}$ wavelets of the filterbank $\psi_j^{(i)}$ with dilating the mother wavelet by means of scaling factors $\lambda_j = 2^{j/Q}$ such as:

$$\psi_j^{(i)}(t) = \lambda_j \psi(\lambda_j t), \quad j \in \{0, \dots, J^{(i)}Q^{(i)} - 1\}$$

For any signal $x(t)$, the first convolutional layer (conv 1 in Figure 17) and the first order scattering coefficients are respectively:

$$U_j^{(1)}(t) = |x * \psi_j^{(1)}| \tag{180}$$

$$S_j^{(1)}(t) = U_j^{(1)} * \phi(t) \tag{181}$$

where ϕ is the average pooling operator (pool 1 in Figure 17), it can be interpreted as low pass filtering with down-sampling to avoid aliasing. This allows for observing larger and larger timescales in the structure of the input signal. The scattering coefficients, obtained at each channel $c \leqslant d$ and at each layer $^{(i)}$ are concatenated yielding the feature array $\mathbf{S} = \{S_j^{(i)}\}_{i \leqslant m, 0 \leqslant j \leqslant J^{(i)}Q^{(i)}-1}$.

These features are used for unsupervised seismic classification of signals, in this work [129], Principal Component Analysis (PCA) was used to reduce the dimensionality of the scattering coefficients, and a Gaussian Mixture Model was used for clustering them in the latent space. The main novelties of this methodology, is that the mother wavelet is retrained to jointly minimize the negative log-likelihood of the clustering and a reconstruction loss of the input signal. Instead of learning all the coefficients of the filterbank of each layer $^{(i)}$, they are obtained by interpolating the mother wavelet in the temporal domain with Hermite cubic splines, and dilating it over the total number of filters $J^{(i)}Q^{(i)}$.

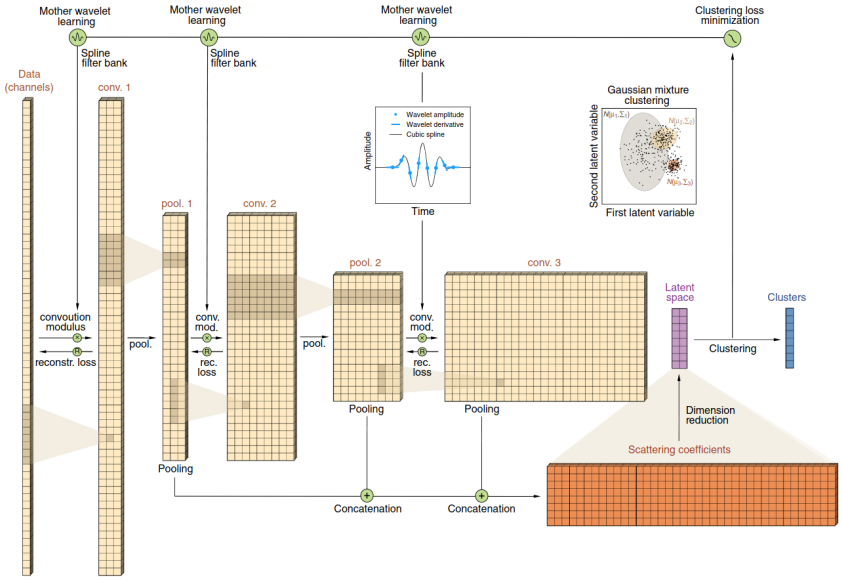


Figure 17: Image taken from [129]

In [133], Maximum pooling was used instead of average pooling $\phi(t)$ and Independent Component Analysis (ICA) [32] was used on the high-dimensional scattering coefficients, followed by a hierarchical clustering for more explainability. Both methodologies were tested on real data and were able to detect precursory signal of Greenland's landslide [129], and seismic signals in a two-day-long seismogram collected in the North Anatolian Fault, Turkey[133].

OTHER APPLICATIONS

In the detection, the primary goal is to minimize the false negative and false positive rates; a similar problem but with a different objective is phase picking, here the focus is on increasing the temporal accuracy of arrival-time picks. This is due to the extreme sensitivity of earthquake location to earthquake arrival time estimates. Authors in [145] propose an architecture based on convolutional networks, that takes as input a three-channel signal $\mathbf{x}(n) \in \mathbb{R}^d$, and outputs two probabilities P_p, P_s of picking respectively a P, and S wave. In [107], authors have proposed a complex architecture that performs simultaneous detection and phase picking.

6.1.3 BEAMFORMING OR MIGRATION TECHNIQUES

Due to the deployment of densely sampled local seismograph arrays, the backprojection or migration technique has become a practical

method for detection of low-magnitude earthquakes [13]. Because of the complex nature of real-data, i. e. low signal to noise ratio, large noise bursts and varying signals polarity, prior to stacking the waveforms are pre-processed.

Consider now we have a network of three seismic sensors (orange triangles in Figure 18). One could directly work on the raw data from these stations, however in practice, data undergoes pre-processing in order to avoid the problem of polarity by producing a positive signal $f(x)$ e. g. envelope was used in [13], and in [88], authors use the kurtosis gradient; prior to aligning and stacking the traces. [88].We then discretize the volume beneath the study region into a grid of k points, each of which representing a possible location of the seismic source. A velocity model associates each potential source k with a collection of P and S wave travel times to each station $\tau^k_{s,c}$.

The characteristic functions applied on each seismic traces of the stations are migrated and stacked according to the velocity model (green trace in Figure18.B).

$$NR_k(t) = \sum_{s,c} f(x_c(t - \tau^k_{s,c})) \tag{182}$$

s, c are the station and channel indexes. In the third and final step, we detect and simultaneously locate the seismic events by analyzing the local maxima of the migration stacks:

$$CNR_{k^*}(t) = \max_k\{NR_k(t)\} = NR_{k^*}(t) \tag{183}$$

Our concern being low-magnitude earthquake detection, the peaks of $CNR_{k^*}(t)$ that surpass a user-defined threshold are events detection located at the source k^*.

The first limitation of this methodology is the need for a velocity model. Then the need to constrain the area underneath the sensors, otherwise the computational complexity gets expensive. The function f should also be smooth enough to allow for seismic traces to stack constructively. Luckily, based on the findings of [88], the kurtosis has two advantages: It disables destructive interference by the misaligned negative phase of the waveform [16] and reduces sensitivity to the velocity model, and compared to the envelope or STA/LTA, backprojection of kurtosis waveforms was the most robust at detecting the smallest-magnitude events [16]. Therefore, we propose a methodology where the sequential detector workflow and this beamforming strategy can be combined in a favourable way.

For evaluation of the different methodologies, we will each time introduce the data on which they were examined, detail the implementation strategies and summarize the detection results.

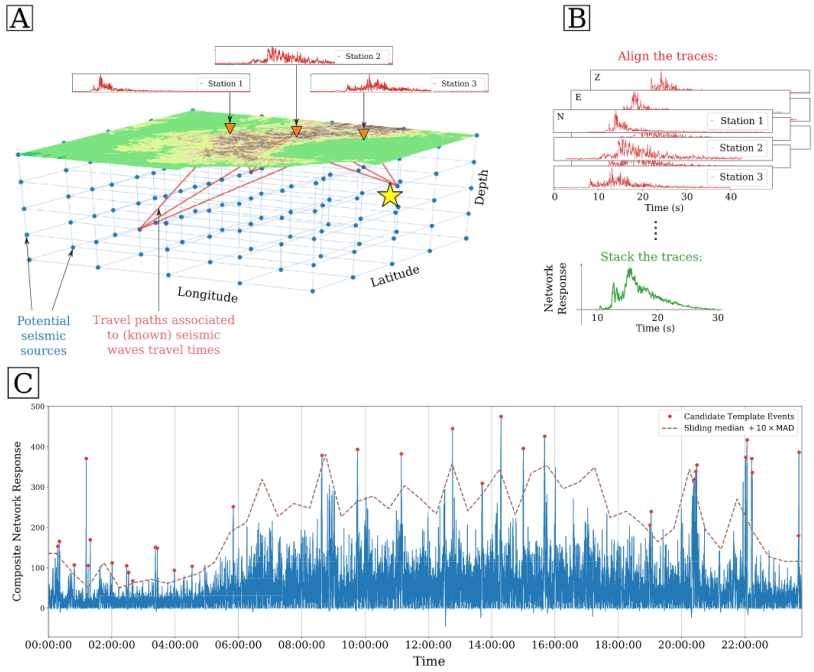


Figure 18: Image taken from [13]

6.2 EVALUATION ON REAL DATA

6.2.1 PRECURSORS TO THE NUUGAATSIAQ LANDSLIDE

On June 17 2017, a landslide occurred in the village of Nuugaatsiaq (northwest of Greenland). This mass slipped into a fjord and generated a destructive tsunami responsible for four fatalities. The landslide and the tsunami were recorded by many stations around the world, we choose to inspect the seismic measurement station in Nuugaatsiaq since it is 30km away from the landslide's location.

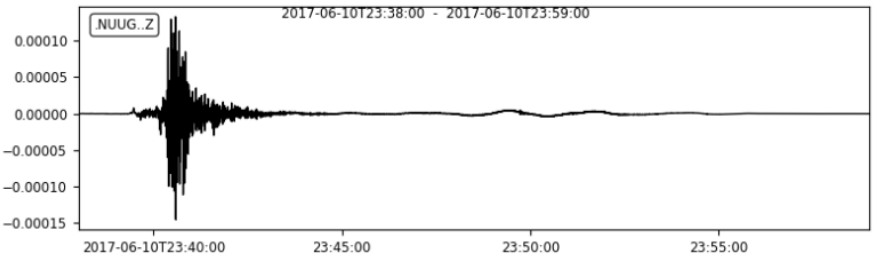


Figure 19: Landslide and tsunami (recorded on the vertical component) occur at 23 : 39.

We will use seismic data from the station NUUG, sampled originally at 30Hz, to detect other precursory signals hidden in the background noise. We select the daylong three-component seismograms

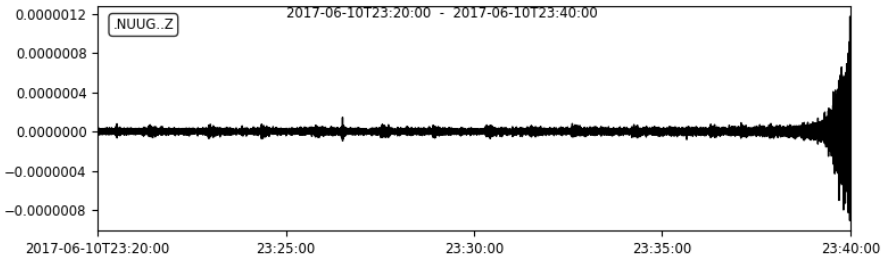


Figure 20: Precursory's signals prior to the mainshock put to evidence by filtering between 2 and 9Hz at t close to the mainshock

from June 17, 2017 00:00 to June 17, 2017 23:38 in order to disregard the mainshock signal (at 23:39) and focus on seismic data recorded before.

CATALOG BY TEMPLATE MATCHING

Template matching exploits the idea of *similarity* between events when the corresponding events occurred within very close proximity of each other. The most effective method of detecting a *known* signal in a potentially noisy time-series is to cross-correlate a waveform template with successive time segments of incoming data. The segments sharing similarity with the template will result in a high value of the correlation matrix. Consider the reference $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T$ and an arbitrary segment of the incoming data $\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T$. The cross-correlation is measured as:

$$\rho = \frac{\sum_{i=1}^{N} x(i)y(i)}{\left(\sum_{i=1}^{N} x(i)^2\right)^{1/2}\left(\sum_{i=1}^{N} y(i)^2\right)^{1/2}}$$

it lies in the interval $[-1, 1]$ with the extreme values occurring only when the two time-series are co-linear.

However, this method evidently requires the knowledge or selection of the reference. To that end, authors in [116] have selected an arbitrary template from the signals very close in time to the mainshock as illustrated in Figure 20.

They correlate it against the day-long seismic data. The resulting correlation trace is compared to a manually-chosen threshold and a potentially precursory signal is detected when it exceeds it. This methodology has yielded 83 newly detected events; there is a clear exponential-like growth of the temporally accumulated event count up to the time of the main shock. The amplitude evolution also follows this exponential-like growth. This behavior agrees with the nucleation model and the results from numerical and laboratory experiments[116].

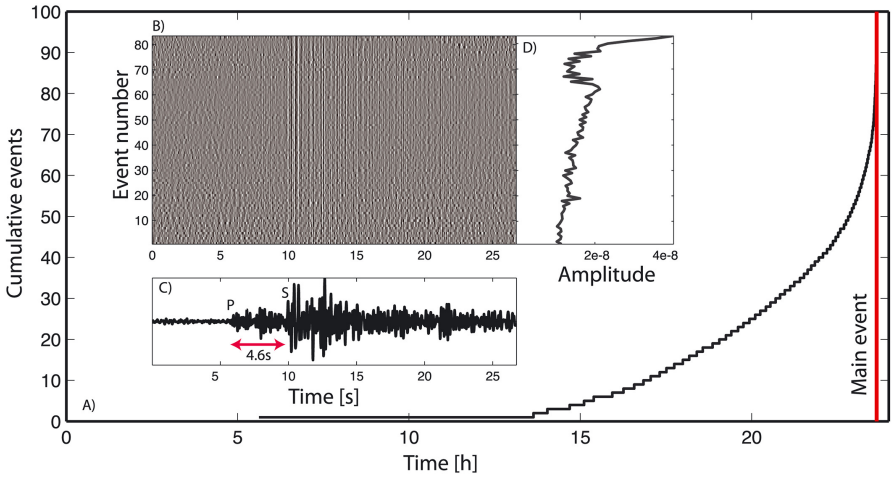


Figure 21: Time evolution of precursory signals. From [116]

Still in practice, rare are the events for which a template reference is known a priori. The template matching method is very sensitive to the quality, the frequency band and time duration of the reference; limiting its use in a "blind" detection problem where few a priori, if not none, information is available about the source. In the following, we explore the use of statistical methods to detect precursory events in a less supervised fashion than the matched detector.

We propose to relate the detection task with the time-series forecasting framework introduced in Chapter 2; we will see how we can recover the precursory signals by means of modeling the ambient seismic noise, the first attempt uses a linear regression model, namely the VAR(p) model, see Figure 23, and in a second stage, our kurtosis test is applied on the residuals of the linear regression, see Figure 24. Aware that the constraint of linearity can be prohibitive, we also implement a recurrent neural network to model the ambient noise. The choice of the LSTM units is justified by their capacity of learning both long and short term dependencies, and from a practical viewpoint, for its robustness against the numerical instabilities of the simple RNN as discussed in Chapter 2.

REVEALING PRECURSORY SIGNALS

The idea is rather intuitive, since we have no a priori on events, we try to model instead the ambient noise for which we have more instances than the rare precursory signals. We split the raw time-series such that we have a training data of noise-only instances on which the model tries to *learn* the regularities in the ambient noise. In the testing phase, we monitor the forecasting errors of the model; if they exceed

a certain threshold then a potential precursory signal is detected. We split the day-long data into a training set from 00:00am to 03:32am (15% of the data), and a test set (03:32am to 23:38pm) stopping exactly before the occurrence of the main shock.

6.2.1.1 *Training a linear VAR(p)*

As a matter of fact, this splitting strategy underlies a very important assumption about the signals: It implies that the ambient seismic noise is a stationary process which is a very optimistic assumption in seismological data. But for now, we assume for now that the ambient seismic noise is a stationary multidimensional auto-regressive VAR(p) process; i. e. satisfying the equation 175. Loosely said, this model estimates the output $\hat{x}_j(t)$ for a channel j, from its own p past values and also the values of the other channels nearby. This is important because we fully exploit the information recorded on multidimensional sensors.

First, the order of the model p was chosen using the Bayesian information criterion (BIC) introduced in Chapter 2, its minimization yielded a value of $\hat{p} = 25$; see Figure 22. The linear auto-regressive

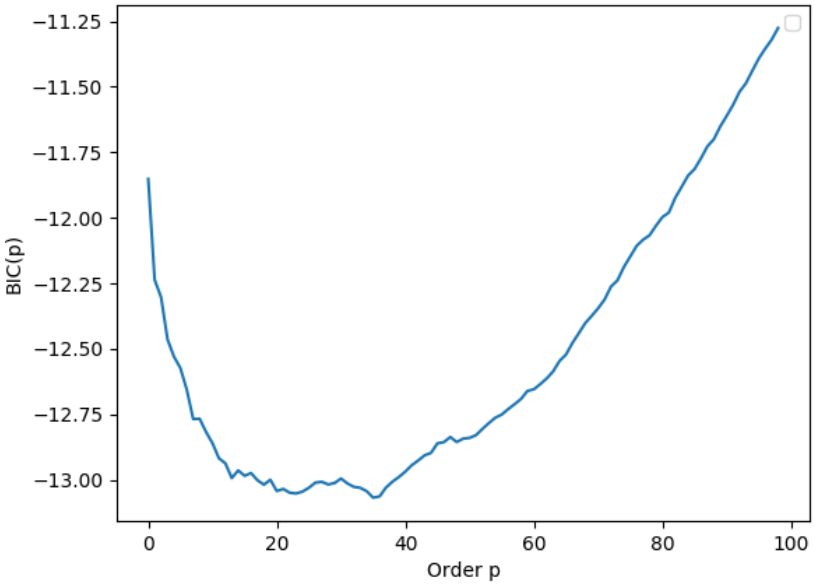


Figure 22: BIC(p) for $1 \leqslant p \leqslant 100$

model's parameters $VAR(25)$ are then estimated from the first 15% of the three-channel daylong data using the least squares method. In the testing phase, we predict the one-step ahead forecast $\hat{\mathbf{x}}(t)$ from the p

previous observations preceding it in the test set. For each time-step, we compute the error

$$e(t) = (\hat{x}(t) - x(t))^T \hat{\Sigma}^{-1} (\hat{x}(t) - x(t))$$

. Wherein $\hat{\Sigma} = \frac{1}{N} \sum_t (\hat{x}(t) - x(t)) (\hat{x}(t) - x(t))^T$ was estimated on the training data.

The evolution of $e(t)$ w.r.t to the number of samples in the test set, is shown in Figure 23. We can see that the error function is peaked

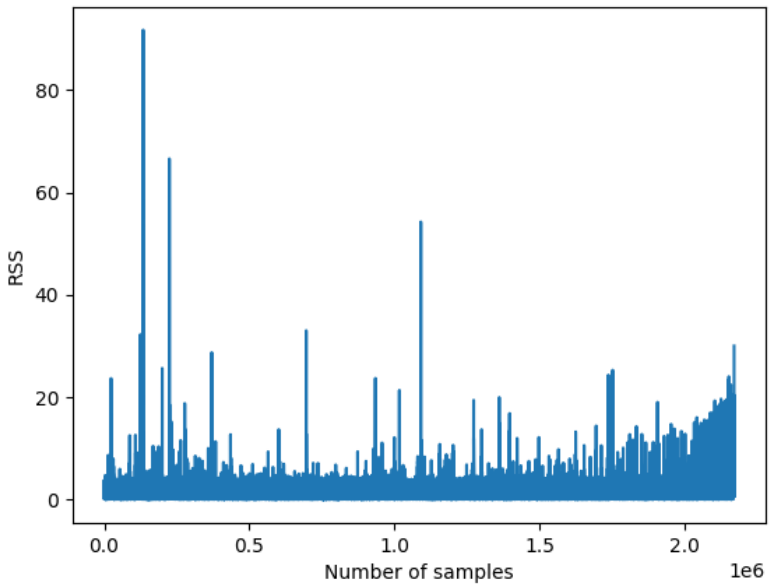


Figure 23: A detector based on second order moments (RSS): High value signals a possible detection of a precursory signal to Greenland's landslide. Acceleration before the mainshock matches the physical model in [116].

at certain times, signaling a big departure of the signal from the expected value of the model. Also, there is an exponential-like growth of detected events right before the mainshock agreeing with the nucleation model discussed in the previous section.

Now remains the important problem of choosing a threshold. Indeed, we can manually choose a threshold by visualizing the evolution of errors, however this empirical approach lacks theoretical guarantees about the false negatives and the false positives of the detection. For this reason, we go beyond simply monitoring the second-order moment of errors in a linear regression problem, and we take advantage of the theoretical guarantees provided by using a stationary VAR(p) model, to frame the detection task in a more robust statistical framework:

Instead of computing $\boldsymbol{e}(t)$, we arbitrarily project the three-channel residuals on a plane, and we compute recursively $\hat{B}_2(t)$ on the now 2D residuals as follows:

$$\hat{B}_2(t) = \lambda_2 \hat{B}_2(t) + (1-\lambda_2)\left(\boldsymbol{\epsilon}(t)^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\epsilon}(t)^T\right)^2$$

where $\boldsymbol{\epsilon}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ and $\lambda_2 = 0.998$ is a fading factor that smooths out the detection characteristic function. To give more intuition, as explained in Chapter 5, this is equivalent to choosing a sliding window of duration 33.3 seconds. The detection characteristic function is shown in Figure 24. Compared to the first strategy based on

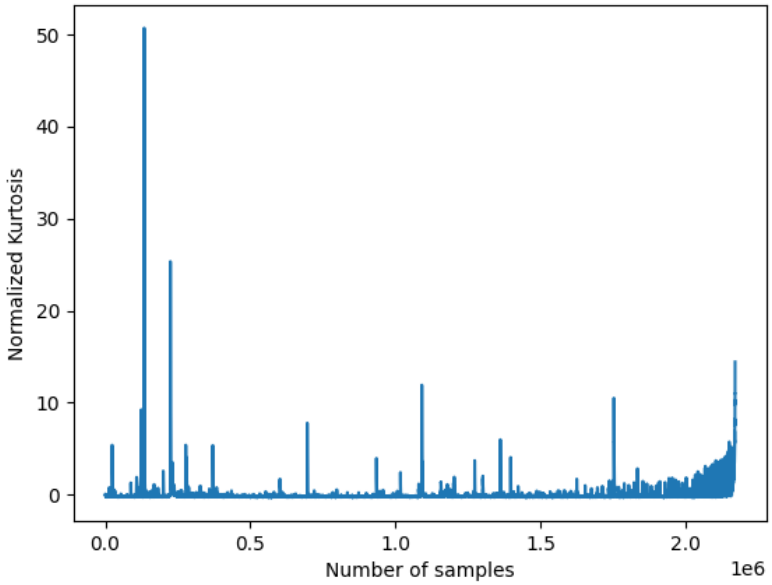
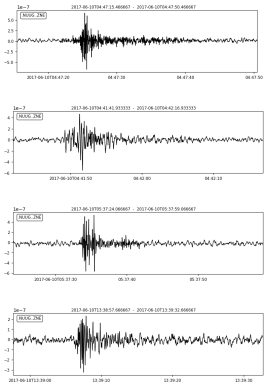


Figure 24: A detector based on the bivariate kurtosis applied to the residuals: High value signals a possible detection of a precursory signal to Greenland's landslide. Acceleration before the mainshock matches the physical model in [116].

the quadratic error vector, the obtained characteristic function has a smoother background and more visible peaks for the onset of precursory signals. The accelerating behavior of the repetitive seismic signals is also visible before the mainshock. By using the results from chapter 4, if the residuals are drawn from a Gaussian distribution, then the standardized $\hat{B}_2(t)$ follows asymptotically a standard normal. We fix the false alarm rate at 5% by using a threshold of ± 1.96.

Results



We obtain 141 newly detected events (four arbitrary examples are plotted on the right). We compare the results of detection obtained by our strategy with the catalog obtained by template matching. To summarize the results, we create an array of all the obtained events and compute a similarity matrix, where each value is the pairwise correlation coefficient between two given seismic signals.

We stress the fact that the goal is not to conclude that we outperform or not the template matching detector. If the reference is known, the latter is optimal for detecting co-located seismic events. The goal is to simply synthesize our results, measuring the overlap with the template matching technique. Just by inspecting 25, we can see that

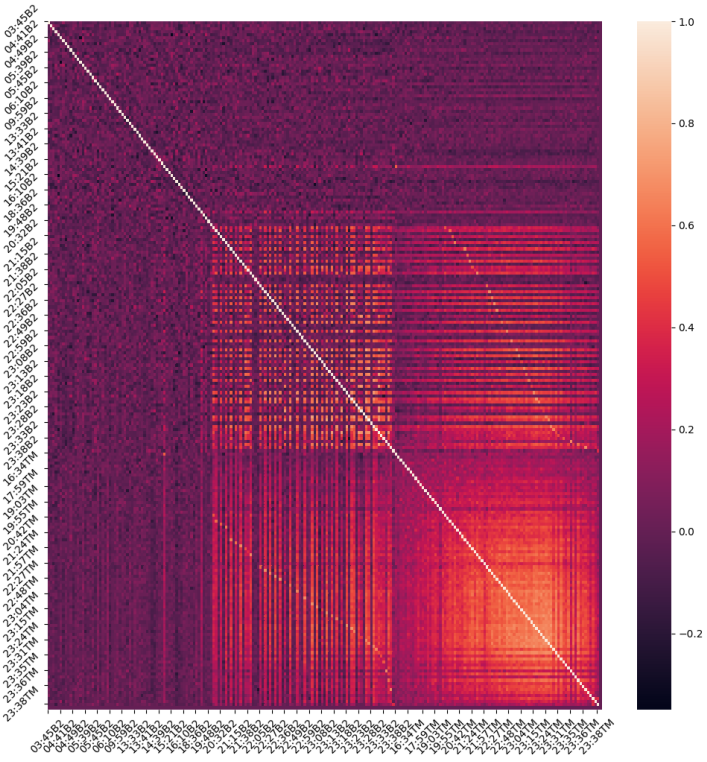


Figure 25: Similarity matrix between all the events. The first 141 rows and columns of the matrix are the events detected by our kurtosis-based strategy, and the last 83 rows and columns correspond to the signals obtained by template matching. The similarity measure is cross-correlation.

we have two distinct classes of signals. The events that were detected long before the main shock starting 01:00am and until 15:00pm are

characterized by a sharp small-duration envelope. These events do not show a strong similarity in time between each other, and they also differ from the events in the catalog provided by template matching. On the other hand, the events detected after 15:00pm, and especially the ones preceding the main shock starting 20:32pm share similarities in time between each other, and also with the other events revealed by template matching.

The fact that the temporal similarity of all these events is particularly visible for later events is probably due to the fact that the signal-to-noise ratio of these events increases toward the landslide. And perhaps, this is what hinders the visibility of similarity between the events occurring before 15:00pm.

To sum up, we have been able to detect 141 of the precursory signals to Greenland's landslide without the need to a reference template; we were able to recover 70 of the events already present in the template matching catalog (84% overlap) and 71 newly detected events. Indeed, tuning the detection threshold by allowing higher or lower false alarm rates will yield a different count of detection. The main feature to keep in mind is that this strategy has theoretical guarantees and allows to fix the false alarm rate. Aware that real-data is complex in nature, with both linear and non-linear components, we study the impact of replacing the VAR(p) model with a Recurrent neural architecture (LSTM) as a candidate model for the ambient noise. These models promise the flexibility to learn both long-term and short-term relationships in the incoming data; but we loose theoretical guarantees about the distribution of the residuals. Thus, there is no justification for using the kurtosis-based normality test. We simply review the training procedure of these complex architectures and monitor the error of the model's forecasts. This study was directed during a 6-month internship of student Louis Closson.

Training LSTM network

When dealing with Neural networks, the first task is is to choose an appropriate architecture: How many layers? How many hidden units per layer? As there is not turn-the-crank procedure and very few theoretical studies, we have adopted the trial-and-error exercise.

Similar to the training procedure of the linear model, the model's parameters (weights and biases) are estimated on the first 15% of the three-channel daylong data using the stochastic gradient descent; the batch size is $M = 256$. The final adopted architecture consists of a two-layer LSTM with 18 hidden units on each layer. Finally, a fully connected layer whose choice is dictated by the dimension of the out-

put ($d = 3$) provides the output. The model learns to predict the output $\hat{\mathbf{x}}(t)$ from 16 previous time-steps $\mathbf{x}(t-16), \ldots, \mathbf{x}(t-1)$. We monitor the evolution of $e(t) = (\hat{\mathbf{x}}(t) - \mathbf{x}(t))^T (\hat{\mathbf{x}}(t) - \mathbf{x}(t))$ w.r.t samples in the test set.

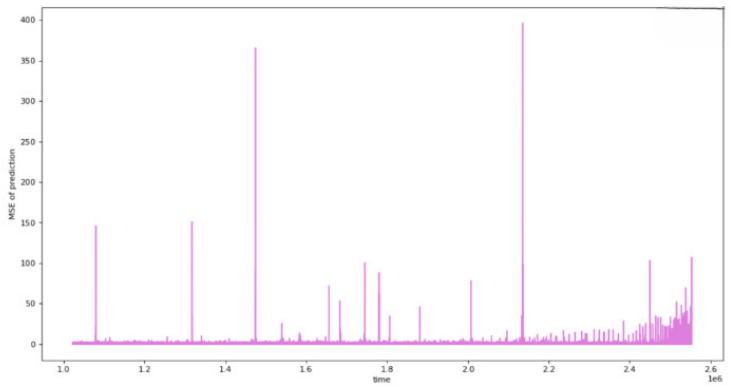


Figure 26: Forecast error e(t) w.r.t to time. Acceleration before the mainshock matches the physical model in [116].

We manually choose a threshold equal to two times the standard deviation of $\boldsymbol{e}$, we obtain 71 newly detected events of which 40 overlap with the template matching catalog (51% overlap) and 31 newly detected events occurring between 3am and 13pm. The similarity matrix summarizing the detection results is in Figure 27.

To conclude this series of simulations, the two-stage detection strategy based on pre-whitening and using the kurtosis-based normality test on the residuals yields encouraging results. For this dataset, increasing the complexity of the model by using LSTM blocks also revealed similar events to the ones recovered by the first strategy, but at the cost of losing the theoreticial guarantees about the false alarm rate. Motivated by these results, we will now put in practice the sequential detector strategy presented in Chapter 5 on a more complex dataset. We will first validate it on one station, and then extend it to a network of stations.

THE NORTH ANATOLIAN FAULT, TURKEY

We also dispose of continuous three-component seismic data from 8 stations of the DANA experiment in Turkey. We choose the data set for mainly two reasons. First of all, the data set contains both seismic and anthropogenic activity, which is a typical situation in most seismological studies[133]. Second of all, an existing template matching catalog provides labels for the seismicity in this area. The catalog was built following the methodology in [13], and for comparison we conduct the same pre-processing steps: The recodings on all stations are

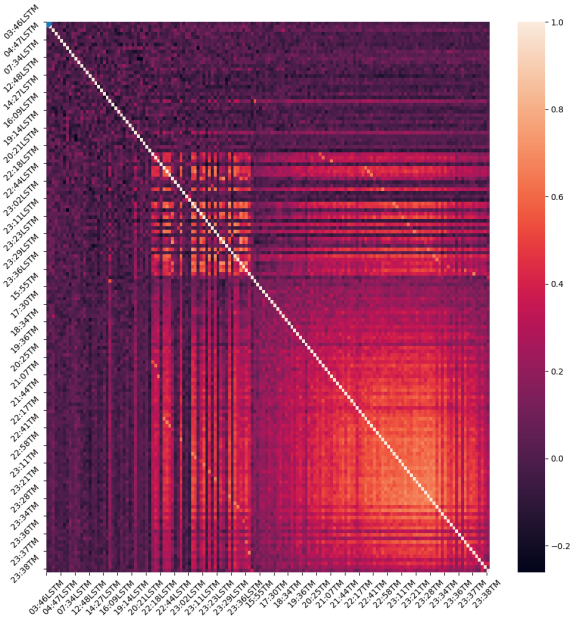


Figure 27: Similarity matrix between all the events. The first 141 rows and columns of the matrix are the events detected by LSTM forecast errors, and the last 83 rows and columns correspond to the signals obtained by template matching. The similarity measure is cross-correlation.

down-sampled to 25Hz and filtered in the band $2 - 12$Hz using a Butterworth band-pass filter. Prior to down-sampling a type 2 Chebyshev low pass filter of order 10 is applied to alleviate aliasing. We choose to analyze the seismic data from the 7th to the 8th of July 2012.

6.2.2 VALIDATION ON ONE STATION OF THE DANA ARRAY

We first start with validating the proposed methodology in Chapter 5, that is pre-whitening, arbitrary bivariate projection and recursive computation of the bivariate kurtosis $\hat{B}_2(t)$, on one station of the DANA array called SAUV. For comparison with a power-based detector, we run a recursive STA/LTA algorithm on one channel of the long-day incoming data, the results are in Figures 29, 30.

The sequential detector has a smoother background than STA/LTA for ambient noise and pronounced peaks signaling the onset of potential seismic events. The power detector STA/LTA on the other hand has many large bursts. This is normal as it is sensitive to the low SNR of the data to hand.

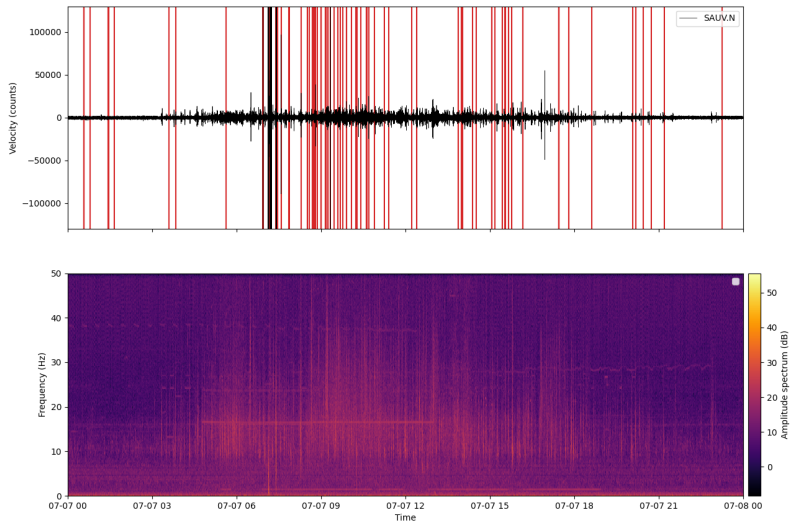


Figure 28: East component of one station SAUV from the DANA array. In red vertical lines, the detected seismic events present in the catalog provided by a similar methodology to [13] used on multiple stations of the DANA array.

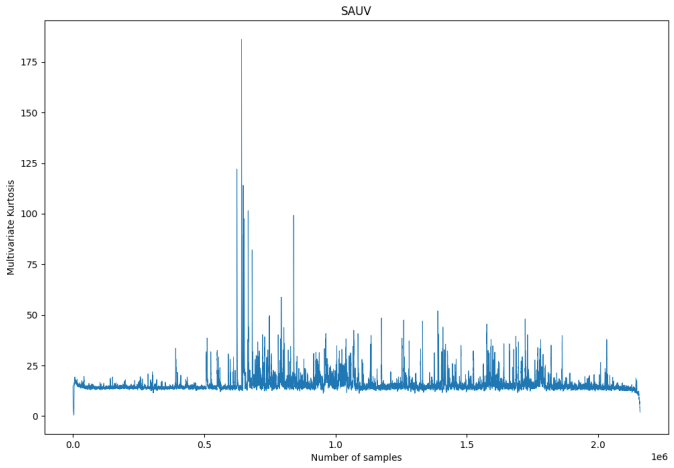


Figure 29: $\hat{B}_2$ w.r.t time time

6.2.3 VALIDATION ON MULTIPLE STATIONS OF THE DANA ARRAY

Even greater improvement in detecting low-magnitude signals can be achieved using a network of stations. The delay-and-sum (beamforming) of traces from closely spaced sensors increases the SNR through a simultaneous summation of coherent signal and cancellation of incoherent noise. We propose the standardized bivariate kurtosis obtained

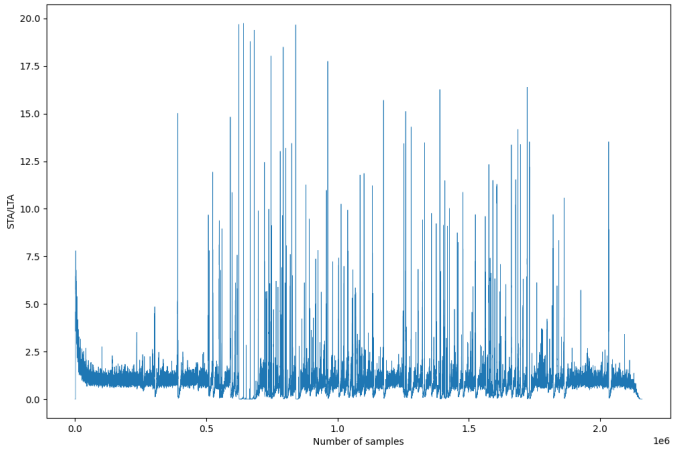
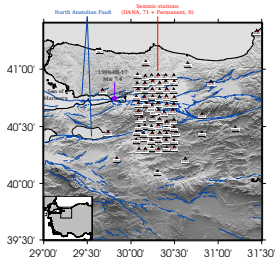


Figure 30: Recursive STA/LTA wherein $N_s = 100$, $N_l = 2000$ (in number of samples)



by recursive pre-whitening as a candidate for the summation. 8 stations

$$\{SAUV, SPNC, DC08, DC07, DC06, DD06, DE07, DE08\}$$

next to SAUV were selected to test the kurtosis as a candidate function for the beamforming method. The data from all stations are downsampled to 25 Hz and filtered in the band 2–12 Hz.

We have discretized the area underneath 8 stations, yielding k potential sources, and we recall that each network's response is obtained by:

$$NR_k(t) = \sum_{s,c} f(t - \tau^k_{s,c}) \tag{184}$$

$$CNR_{k^*}(t) = \max_k NR_k(t) \tag{185}$$

Where the moveouts $\tau^k_{s,c}$ were computed using the ray-tracing software Pykonal in the 1D velocity model due to [79]. The goal is to compare three choices of $f(.)$: the envelope of the traces, the univariate kurtosis on one arbitrary axe and finally the workflow proposed in Chapter 5, that is for each station, seismograms are pre-whitened using multidimensional autoregressive filtering. The obtained residuals will then be arbitrarily projected on a plane on which the bivariate kurtosis is estimated $\hat{B}_2$. In Figures 31, 32, 33, the Composite Network response of the envelope, the univariate kurtosis and the bivariate kurtosis obtained by our sequential detector are shown w.r.t time. Remains the important question of choosing the threshold. We can no

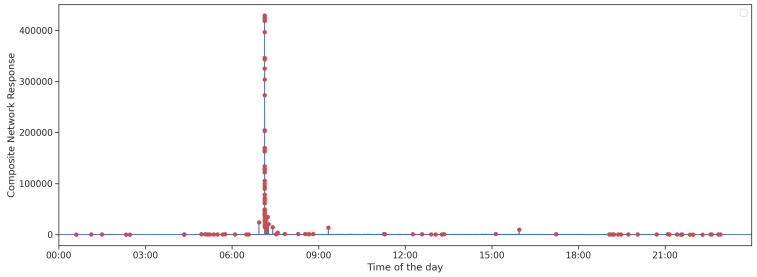


Figure 31: $\mathrm{CNR}_{k^*}(t)$ computed using $\mathrm{NR}_k(t) = \sum_{s,c} env(t - \tau_k^{s,c})$ where f is the envelope of the seismic trace. Red peak corresponds to a possible detection.

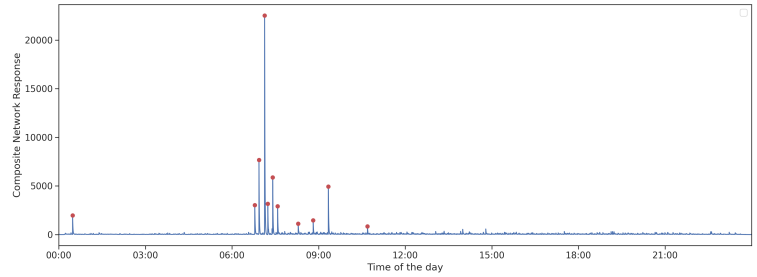


Figure 32: $\mathrm{CNR}_{k^*}(t)$ computed using one channel $\mathrm{NR}_k(t) = \sum_s \hat{B}_1(t - \tau_k^s)$ computed after pre-whitening using multi-dimensional auto-regressive filtering

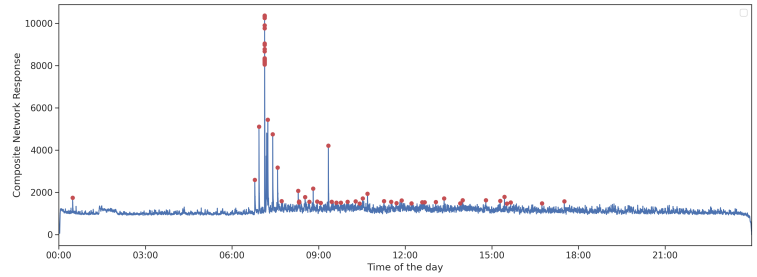


Figure 33: $\mathrm{CNR}_{k^*}(t)$ computed using the $\mathrm{NR}_k(t) = \sum_s \hat{B}_2(t - \tau_k^s)$ computed after pre-whitening using multi-dimensional auto-regressive filtering

longer use the critical values of a standard normal because the composite network response is the maximum of the sum of $\hat{B}_d$ ($d = 1, 2$). Under the hypothesis of Gaussianity, we know that asymptotically $\sum \hat{B}_d(t)$ is also Gaussian with mean and variance detailed in Chapter 4. Hence, we can use the results on the maximum of random Gaussian variables to derive the distribution of the CNR. For now, we manually pick the threshold by visualizing the histogram of the CNR.

The composite network response obtained by the envelope detector favors signals with great energy, hence we observe a dominant peak corresponding to the event with the highest magnitude, drowning all the rest of the events. Therefore authors in [13] propose the use of an adaptive threshold to find maxima locally in a sliding window. Using our sequential detector with univariate kurtosis already yields better results compared to stacking the energy but the count of events is smaller compared to using the bivariate kurtosis on an arbitrary projection. The scalar kurtosis revealed 23 seismic events, whereas the bivariate kurtosis revealed 98.

6.3 CONCLUSION AND CONTRIBUTIONS

The ever increasing amount of geophysical data at our disposal gives us unprecedented possibilities to devise sophisticated models; but it also requires careful efforts in designing statistically significant and efficient methods from theoretical and implementation viewpoints. For this reason, we have proposed an operational sequential detector to skim through the incoming data and reveal the onset of seismic events. The two-stage workflow consists of filtering the data using a linear vector auto-regressive model to capture the linear component of the data. If the sequence contains mostly noise, the residuals are expected to be Gaussian justifying the use in a second stage of the kurtosis-based normality test to detect the onset of seismic events. These events are expected to be sharp peaked, which translates as a heavy-tailed distribution with a high kurtosis. We also propose to reduce the dimension of the initial residuals by projecting them on a plane, in order to use the limiting distribution derived in Chapter 4.

In the proposed workflow, only few parameters are to be tuned: The fading factors in the recursive least squares estimation and the recursive estimation of the standardized kurtosis and finally the false alarm rate. The choice of the first two is constrained by the time-scale of changes relative to each data and the last one depends on the desired trade-off between false alarm rate and power of detection. The theoretical guarantees free us from the manual selection of an empirical threshold. This workflow has also the advantage of not requiring large labeled data to train the model on, it adapatively learns the parameters of VAR(p) using an efficient implementation of the RLS algorithm. It also allows the model to be more flexible to changes in the background noise and generalizes well to different datasets.

Our strategy was able to recover events detected by the template matching strategy without requiring a reference waveform and also reveal newly detected events. It has also been extended to seismic

arrays and the beamforming based on the bivariate kurtosis yielded better results than energy-based beamforming techniques. Our exploratory detection task using LSTM on the Greenland's landslide has demonstrated that increasing the complexity of the model does not necessarily lead to better detection results. With no intention of being conclusive based on one set of data, but this finding demonstrates that the ever increasing amount of data should not always imply the implementation of an equally complex model. On the contrary, scalable methods with theoretical guarantees should be preferred to allow for more generalization, explainability and efficiency.

CONCLUSION

In this manuscript, we investigated detection techniques adapted to d-dimensional time-series, with an application to seismology. We gave a great deal of attention for designing an efficient detector that can run with low computational burden on large datasets, and with theoretical guarantees on the false alarm rate.

How to derive a detector with theoretical guarantees on the false alarm rate?

In this work, we tackled the detection task in a statistical framework, where detecting changes in the distribution of the time-series amounts to the departure of a test statistic from Gaussianity. The main result of our work presented in Chapter 4 derived the limiting distribution of Mardia's measure of Kurtosis on bivariate time-series. Subsequently, we generalized these findings to d-variate data by means of random projections. Supported by numerical results on colored copula, we have shown that testing normality with bivariate kurtosis on 2D-projections outperforms its scalar counterpart on 1D-projections. Our work strongly motivates taking into account both the spatial and temporal dependence when testing the Gaussianity of time-series.

How to translate the theoretical findings to an operational detector?

The testing procedure should now run with low computation burden on large datasets and reveal non-Gaussien signals embedded in Gaussian noise. To that end, we proposed in Chapter 5 a two-stage sequential detector prior to computing the test statistic recursively, time-series are pre-whitened using VAR(p) model, or even randomly projected if their dimension $d \geqslant 3$; in which case the false discovery rate is controlled using a Benjamini-Hocheberg procedure. In the absence of events, the null hypothesis is that residuals are a realization of a Gaussian process. This workflow yields good detection power results on both synthetic and real-word data. This is supported by numerical experiments in Chapter 5 and comparisons with energy-based detectors, and an anomaly detector based on ANN in Chapter 6.

We were tempted to believe that increasing the complexity of the model by using LSTM in Chapter 6 on large data would increase the detection power, but it yielded similar performances to our sequential detector. Why were we tempted to believe that in first place is the first open question we raise in the following.

Rather than deriving the calculations manually, a natural continuation of this work would be to implement the calculation steps in a symbolic computing tool; or perhaps derive bounds on the mean and variance of the test statistic for higher-dimensional data.

OPEN QUESTIONS

The hype around ANN and the unprecedented results they obtain on complex tasks has made them an ubiquitous tool. It seems that for a learning method to be admissible, it *should* contain a flavour of ANN. We have chosen to introduce ANN in the statistical framework of Chapter 2. This is coherent because they are in the end simply non-linear statistical models. We have also raised concerns related to the choice of their architecture, training issues and lack of generalization and explainability. Another important issue is that training over-parameterized networks requires significant computing costs, and by extension energy costs. These important issues are, if addressed, supposedly alleviated by empirical methods (with a strong judgmental component) such as pruning or early stopping.

This a heavy list of heavy disadvantages somehow goes unbeknownst to the unaware practitioner, who with trial-and-error, obtains the desired performance. We do not attempt to pass judgment on the merits of these methods, on the contrary, we encourage their use when that is necessary and appropriate to the problem at hand. In deciding that, one should not be disconnected from the global cost and implications of this necessity.

Recently, these questions are more addressed in research works, for example considerable attention is put on formulating a mathematical framework for ML practices, elaborating on approaches like the Probably Approximately Correct learning [138] and related approaches The introduction of PAC has done an admirable job of drawing together ML practitioners with computer scientists to seek answers to *How likely is a learner to output an approximately correct model?*

The intent of PAC learning is that successful learning of unknown targets entails obtaining, with high probability, a hypothesis that is a good approximation of the the target. Formally, the basic model assumes that instances are in $\{0,1\}^n$, but it can easily be extended to non-Boolean based attribute instance spaces. A *concept* is PAC learnable by a set of hypotheses (or models) $\mathbb{H}$ if there is a polynomial time learning algorithm and a polynomial $p(n)$ such that for $n \geqslant 1$ and $\epsilon > 0$, $0 < \delta < 1$, if the algorithm A is given at least $p(n)$ independent samples of the concept targets, then with probability at least $1 - \delta$, A returns a hypothesis h such that error$(h) \leqslant \epsilon$. $p(n)$ is the *sample complexity* of the learning algorithm A.

How much data one must see to satisfy a specific pair (δ, ϵ) depends on how complex the given class of hypotheses are. By increasing the size of the hypothesis space, it may become easier to find a good approximation, but that requires passing through more train-

ing samples.. The bias induced by restricting the hypothesis space $\mathbb{H}$ is quantified by using the Vapnik-Chervonenkis dimension denoted $\dim_{VC}(\mathbb{H})$ [19]. It can be shown that the sample complexity of a learning algorithm by $\mathbb{H}$ is bounded by [65]:

Computational complexity vs. sample complexity

However in practice, this bound is considered a loose overestimate.

$$\frac{1}{\sqrt{\epsilon}\sqrt{1-\epsilon}}\left(2\dim_{VC}(\mathbb{H})\ln(\frac{6}{\epsilon})+\ln(\frac{2}{\delta})\right)$$

An active area of reasearch is to extend the PAC theory to neural networks. The goal is to develop analytic tools to help understand the problem of generalization and overfitting in these more complex decision rule spaces. Further details about the theory of PAC can be found in [56, 130]. Another active area of research is concerned with the choice of architecture of ANN. To circumvent the judgemental component of the commonly used methods such as: regularization, early stopping and pruning, attempts have been made to derive statistical hypothesis testing in [5], or derive Information criteria for a feedforward ANNs [109].

We have also discussed the use of hybrid architectures in Chapter 2 and gave an example of merging ARMA with ANN. Another example is the scattering network that combines the favours of the theoretical guarantees of the scattering transform with the learning flexibility of CNN. This combination of expert knowledge with the ANN performance holds the promise of a best-of-both world scenario, in which *fair, accessible* and almost *explainable* outputs are given to the practitioner.

In our work, we could make the results obtained by the sequential detector more reliable and conclusive by *clustering*. This exercise of categorization, where the events are grouped based on a *similarity measure* summarizes information about the multiple detected events. This is a proposal of a hybrid architecture, where our kurtosis-based detector reveals non-Gaussian processes, and then a clustering method, such as spectral clustering, takes the "torch" to reveal the different communities of these complex events.

FUTURE DIRECTIONS APPLICATION DIRECTIONS

Turning real-data into insights is an exciting aspect that holds promises of understanding the dynamics that rule Earth, and perhaps contribute in early warning systems where seismic events are predicted and therefore prevented.

forecasting Earthquakes or volcanic eruptions is still in its infancy [122]

The application of detecting unseen patterns in data has naturally excluded the use of supervised deep learning practices, because of the lack of a large training dataset, and also due to their inability to

provide guarantees on the false alarm rate. This is why we favoured a methodology with more theoretical guarantees based on Mardia's kurtosis estimated on the linear regression model residuals. This is a valid approach (and very common in signal processing), where we supposed that in the absence of a peaked seismic signal, the sensors record small fluctuations of multiple noise source and multiple scattering noises (and thus Gaussian by the Central Limit theorem). Instead of modeling the complex signals of interest, we model what is recorded in their absence to reveal them. We also note that the use of higher order statistics in seismology is not new; the kurtosis is commonly used in phase-picking applications; What is new in our work is choosing a threshold of the bivariate kurtosis with a fixed false alarm rate, which is suitable for multivariate time-series.

Phase-picking

If the problem at hand is now phase picking discussed in Chapter 6 where onset times should be chosen carefully to enable the location of the sources, the proposed sequential detector should be run twice, first on the time-series, and subsequently on its time-reversed counterpart. It is important to note that our test statistic is computed recursively with a fading factor. This causes a lag in the detection of the onset of seismic events. We could re-estimate the test statistic, this time on its time-reversed to yield better estimates of the onset of the events.

A Clustering point of view

Alternatively, the problem at hand could have been framed as clustering this plethora of complex events, in which case turning to more complex models would be required. See for example the work of [108], [129, 133] to name a few. In a similar spirit, exploring architectures such as Autoencoders for where the values of the reduced latent space form an interest set of features for clustering seismic events. This would be an interesting complementary approach to our work.

Kurtosis-based beamforming

More effort should be made on generalizing the kurtosis based beamforming to more periods of data or different regions to derive more general conclusions on this procedure. Additionally, the threshold chosen empirically for now for the composite network response (which is the maximum at each time-step t of the values of $\hat{B}_2$ from multiple stations), should be replaced with a more robust choice. If we can

assume that the network responses of each station are independent Gaussian random variables, then the asymptotic distribution of the maximum (the composite network response) is the standard Gumbel distribution. This reiterates the advantage of using a scalar test statistic $\hat{B}_2$ on a multivariate problem, for which deriving such limiting distributions is possible analytically.

Part IV

APPENDIX

A

APPENDICES

A.1 APPENDICES

A.1.1 PROOFS OF LEMMAS 4.2.1 AND 4.2.2

Proof. of Lemma 4.2.1. Under Hypothesis $\mathcal{H}_0$, the covariance of entries Δ_{ab} take the form below :

$$\mathrm{Cov}(\Delta_{ab},\Delta_{cd}) = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\{x_a(n)x_b(n)x_c(m)x_d(m)\} - S_{ab}S_{cd}$$

and letting $\tau = n-m$, and $\Omega_{abcd} = S_{ac}S_{bd} + S_{ad}S_{bc}$ we have after some manipulation:

$$\begin{aligned}\mathrm{Cov}(\Delta_{ab},\Delta_{cd}) &= \frac{1}{N}\Omega_{abcd} + \frac{2}{N}\sum_{\tau=1}^{N-1}(1-\frac{\tau}{N})\{S_{ac}(\tau)S_{bd}(\tau) + S_{ad}(\tau)S_{bc}(\tau)\} \\ &\leqslant \frac{1}{N}\Omega_{abcd} + \frac{2}{N}\sum_{\tau}\{|S_{ac}(\tau)|\,|S_{bd}(\tau)| + |S_{ad}(\tau)|\,|S_{bc}(\tau)|\}.\end{aligned}$$

Next, using the inequalities $|\sum_i u_i v_i| \leqslant \sum_i |u_i||v_i| \leqslant \frac{1}{2}\sum_i (u_i^2+v_i^2)$, we have:

$$|\mathrm{Cov}(\Delta_{ab},\Delta_{cd})| \leqslant \frac{|\Omega_{abcd}|}{N} + \frac{1}{N}\sum_{\tau}|S_{ac}(\tau)|^2 + |S_{bd}(\tau)|^2 + |S_{ad}(\tau)|^2 + |S_{bc}(\tau)|^2.$$

Now using the mixing condition stated in page 72, $\sum_{\tau=0}^{\infty}|S_{ij}(\tau)|^2 \leqslant \Omega_{ij}$, we eventually obtain:

$$|\mathrm{Cov}(\Delta_{ab},\Delta_{cd})| \leqslant \frac{|\Omega_{abcd}|}{N} + \frac{1}{N}(\Omega_{ac}+\Omega_{bd}+\Omega_{ad}+\Omega_{bc}) \quad (186)$$

which shows that $\mathrm{Cov}(\Delta_{ab},\Delta_{cd}) = O(1/N)$. □

Proof. of Lemma 4.2.2. Notice that positive definite sample covariance matrix may be reexpressed as

$$\hat{\mathbf{S}} = \mathbf{S} + \boldsymbol{\Delta} = \mathbf{S}^{1/2}\mathbf{I}\mathbf{S}^{1/2} + \mathbf{S}^{1/2}\mathbf{S}^{-1/2}\boldsymbol{\Delta}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}$$

Let $\mathbf{E}$ be the symmetric matrix $\mathbf{E} = -\mathbf{S}^{-1/2}\boldsymbol{\Delta}\mathbf{S}^{-1/2}$. Then with this definition,

$$\hat{\mathbf{G}} = \mathbf{S}^{-1/2}(\mathbf{I}+\mathbf{E})^{-1}\mathbf{S}^{-1/2}$$

As for any matrix $\mathbf{E}$ with spectral radius smaller than 1, the series $\sum_{k=0}^{\infty} \mathbf{E}^k$ converges to $(\mathbf{I} - \mathbf{E})^{-1}$. If we plug this series in the expression of $\hat{\mathbf{G}}$, for N large enough to warrant that the spectral radius of $\mathbf{E}$ is less than 1, we get $\hat{\mathbf{G}} = \mathbf{S}^{-1/2} \sum_{k=0}^{K} \mathbf{E}^k \mathbf{S}^{-1/2} + o(\|\mathbf{E}\|^K)$. Replacing $\mathbf{E}$ by its definition and taking $K = 3$ eventually yields (143). Note that the precise approximation order is $O(N^{-3/2})$, but only $o(1/N)$ will be useful in what follows. □

A.1.2 MCCULLAGH'S BRACKET NOTATION AND EXPRESSION OF THE HIGHER MOMENTS UNDER THE NULL HYPOTHESIS

McCullagh's bracket notation [101] allows to write into a compact form a sum of terms that can be deduced from each other by generating all possible partitions of the same type. For instance, we have the following expression for fourth order moments M_{abcd} of a zero-mean multivariate normal variable with covariance $\mathbf{S}$:

$$M_{abcd} = S_{ab}S_{cd} + S_{ac}S_{bd} + S_{ad}S_{bc} = [3]S_{ab}S_{cd} \tag{187}$$

Moments of higher order can be found easily:

$$\text{order 6:} \quad M_{abcdef} = [15]S_{ab}S_{cd}S_{ef} \tag{188}$$

$$\text{order 8:} \quad M_{abcdefgh} = [105]S_{ab}S_{cd}S_{ef}S_{gh} \tag{189}$$

$$\text{order 10:} \quad M_{abcdefghij} = [945]S_{ab}S_{cd}S_{ef}S_{gh}S_{ij} \tag{190}$$

$$\text{order 12:} \quad M_{abcdefghijk\ell} = [10395]S_{ab}S_{cd}S_{ef}S_{gh}S_{ij}S_{k\ell} \tag{191}$$

$$\text{order 14:} \quad M_{abcdefghijk\ell mn} = [135135]S_{ab}S_{cd}S_{ef}S_{gh}S_{ij}S_{k\ell} S_{mn} \tag{192}$$

$$\text{order 16:} \quad M_{abcdefghijk\ell mnpq} = [2027025]S_{ab}S_{cd}S_{ef}S_{gh}S_{ij}S_{k\ell} S_{mn}S_{pq} \tag{193}$$

since it is well known that there are $[\frac{2r!}{2^r\, r!}]$ terms in the moment of order $2r$.

A.1.3 CALCULATION METHODOLOGY

Remind that, as introduced in Lemma 4.2.3, $A_{\alpha^l\beta^l} = \mathbf{x}(\alpha_l)^T G \mathbf{x}(\beta_l)$, where G stands for the true precision matrix of the process whose terms are $G_{r,c}$, and where $(r, c) \in \{1, \ldots, d\}^2$.

Referring to the expression of $\hat{B}_p$ or $\hat{B}_p^2$ as derived from equation (145), it appears that the indices (α_l, β_l) take values on a restricted

set $\mathcal{S} = \{i, j, k, \ldots\}$, and $|\mathcal{S}| \ll N$. The following compact notation is therefore introduced

$$\mu^{\alpha_1 \ldots \alpha_L \beta_1 \ldots \beta_L}_{r_1 \ldots r_L c_1 \ldots c_L} = M_{i^{\eta_i} j^{\eta_j} k^{\eta_k} \ldots} \tag{194}$$

where

$$\eta_i = \sum_{l=1}^{L} \left(\mathbb{I}_{[\alpha_l = i]} + \mathbb{I}_{[\beta_l = i]} \right), \forall i \in \mathcal{S}$$

Note that the subscripts $r_1, \ldots c_1 \ldots$ are skipped here for sake of readability, though any permutation of the superscripts in equation (146) requests the corresponding permutation of the subscripts. It is easier to describe the general methodology by the typical example below.

Example

Consider the moment $\mathbb{E}\{A_{nn} A_{nj} A_{jk} A_{kn}\}$. According to equation (147) it will be expanded as a sum of moments of order 8 (i.e. $L = 4$); using the compact notation from equation (194), we get

$$\begin{aligned} \mathbb{E}\{A_{nn} A_{nj} A_{jk} A_{kn}\} &= \sum_{((r_i, c_i)_{i=1 \ldots 4}) = 1}^{d} G_{r_1 c_1} G_{r_2 c_2} G_{r_3 c_3} G_{r_4 c_4} \mu^{nnnjjkkn}_{r_1 c_1 r_2 c_2 r_3 c_3 r_4 c_4} \\ &= \sum_{((r_i, c_i)_{i=1 \ldots 4}) = 1}^{d} G_{r_1 c_1} G_{r_2 c_2} G_{r_3 c_3} G_{r_4 c_4} M_{n^4 j^2 k^2} \end{aligned} \tag{195}$$

The sum involves $2^{2L} = 64$ terms. It is reminded that the coefficients r_i or c_i indicate the coordinate of the vector process (or space coordinate, thus taking values on $\{1, \ldots, p\}$) , whereas time indices n, j, k tale values on$\{1, \ldots, N\}$. Following McCullagh's notations, under the assumption ($\mathcal{H}_0$) that the d-dimensional process is centered and jointly Gaussian, for this particular 8-th order moment

$$M_{abcdefgh} = [105] S_{ab} S_{cd} S_{ef} S_{gh}$$

which expresses that under $\mathcal{H}_0$, higher even order moments (odd-order moments are zero) may be expanded as sums of products of second order moments. It must be reminded that here, a, b, c, d, e, f, g, h stand for 'meta-indices' defined in the present example by

$$(n, r_1), (n, c_1), (n, r_2), (n, c_4), (j, c_2), (j, r_3), (k, c_3), (k, r_4)$$

respectively, as it appears in equation (195). Plugging the above expansion in equation (195) leads to summing over 64×105 terms! However, in most cases of interest many terms may be grouped together

and highlight the behavior of equation (147). The case $d = 1$ is briefly sketched below as an illustration.

The case $d = 1$ implies that $r_i = c_i = 1\ \forall i \in \{1, \dots, (L = 4)\}$; the particular 8-th order moment in equation (195) may be simply written as $M_{n^4 j^2 k^2}$, whose expansion into sum of products of second order moments will involve the following products : (as there is no ambiguity in this case, we set $M_{ij} \overset{nota.}{=} S_{ij}$),

$$\begin{array}{ll} S_{nn}S_{nn}S_{jj}S_{kk} & \text{appearing 3 times} \\ S_{nn}S_{nn}S_{jk}S_{jk} & \text{appearing 6 times} \\ S_{nn}S_{nj}S_{nj}S_{kk} & \text{appearing 12 times} \\ S_{nk}S_{nk}S_{nj}S_{nj} & \text{appearing 24 times} \\ S_{nj}S_{jk}S_{nk}S_{nn} & \text{appearing 48 times} \\ S_{nn}S_{nk}S_{nk}S_{jj} & \text{appearing 12 times} \end{array}$$

For example the number of occurences of the term of type $S_{nk}S_{nk}S_{nj}S_{nj}$ is given by

$$(4 \times 2 \times 3 \times 1)/2 \times (2 \times 2 \times 1 \times 1)/2 = 24$$

where 4×2 stand for the number of possible choices for index i (one out of 4) times the number of possible choices for index k (one out of 2); then 3×1 stand for the number of remaining possibilities to select index i times the remaining choices for k; Division by 2 accounts for the fact that permutations of terms S_{ik} were counted twice. All other occurence calculations follow the same guidelines. Finally, one gets for the case $d = 1$

$$M_{n^4 jjkk} = 3S_{nn}^2 S_{jj}S_{kk} + 6S_{nn}^2 S_{jk}S_{jk} + 12S_{nn}S_{ij}^2 S_{kk} + 24S_{nk}^2 S_{nj}^2 + \dots. \\ 48S_{nj}S_{jk}S_{nk}S_{nn} + 12S_{nn}S_{nk}^2 S_{jj}$$

which can be directly plugged into equation (195). Note that the sum of all coefficient is actually 105, as expected for an 8-th order moment.

The cases $d \geqslant 2$ turns out to be a bit more complicated, as one has to deal with the 'meta-indices' directly. However counting the number of configurations involving the same time indices follows the same lines as in the case $d = 1$. Going back to the example introduced above for $d = 2$, one gets

$$\mathbb{E}\{A_{nn}A_{nj}A_{jk}A_{kn}\} = \sum_{((r_i,c_i)_{i=1\dots4})=1}^{d} G_{r_1c_1}G_{r_2c_2}G_{r_3c_3}G_{r_4c_4}\ \{\ [3]\ \mu_{r_1c_1}^{nn}\mu_{r_2c_4}^{nn}\mu_{c_2r_3}^{jj}\mu_{c_3r_4}^{kk} \\ + [6]\ \mu_{r_1c_1}^{nn}\mu_{r_2c_4}^{nn}\mu_{c_2c_3}^{jk}\mu_{r_3r_4}^{jk} + [12]\ \mu_{r_1c_1}^{nn}\mu_{r_2c_2}^{nj}\mu_{c_4r_3}^{nj}\mu_{c_3r_4}^{kk} + [24]\ \mu_{r_1c_3}^{nk}\mu_{c_1r_4}^{nk}\mu_{r_2c_2}^{nj}\mu_{c_4r_3}^{nj} \\ + [48]\ \mu_{r_1c_2}^{nj}\mu_{r_3c_3}^{jk}\mu_{c_1r_4}^{nk}\mu_{r_2c_4}^{nn} + [12]\ \mu_{r_1c_1}^{nn}\mu_{r_2c_3}^{nk}\mu_{c_4r_4}^{nk}\mu_{c_2r_3}^{jj}\ \}$$

where we have used notations $\mu_{rc}^{\alpha\beta}$ to emphasize that the permutations (whose number is indicated using McCullagh's brakets) are applied on the 'meta-indices' and grouped such that they share the same 'time structure'; This allow to get the same values as in the case $d = 1$, though replacing the scalar coefficients by McCullagh's brakets.

A.1.4 MULTIVARIATE MOMENTS UP TO ORDER 12

In this section, we give all moments of a zero-mean multivariate normal variable of even order. Most of these expressions have not been reported in the literature. In addition, for the sake of readability, when an index is repeated more than three times, we assume an alternative notation, for instance at order 10:

$$M_{iiiiijjjjk} = M_{i^5j^4k}$$

Furthermore, we use notation introduced in (194) involving meta-indices; more precisely, since each subscript is always associated with a superscript, we may omit the subscript. In order to lighten notation, especially when terms need to be raised to a power, we put the latter superscript in subscript. For instance in (196), M_{abcd}^{iiij} is replaced by M_{iiij}. In the list below, moments are sorted by increasing D, where D denotes the number of distinct indices.

Order 4, D=2.

$$\begin{aligned} M_{iiij} &= [3]\mu_{ab}^{ii}\mu_{cd}^{ij} \\ M_{iijj} &= [2]\mu_{ab}^{ij}\mu_{cd}^{ij} + \mu_{ab}^{ii}\mu_{cd}^{jj} \end{aligned}$$

Order 4, D=3.

$$M_{iijk} = \mu_{ab}^{ii}\mu_{cd}^{jk} + [2]\mu_{ab}^{ij}\mu_{cd}^{ik}$$

Order 6, D=2.

$$\begin{aligned} M_{i^5j} &= [15]\mu_{ab}^{ii}\mu_{cd}^{ii}\mu_{ef}^{ij} \\ M_{i^4jj} &= [12]\mu_{ae}^{ij}\mu_{bf}^{ij}\mu_{cd}^{ii} + [3]\mu_{ab}^{ii}\mu_{cd}^{ii}\mu_{ef}^{jj} \\ M_{iiijjj} &= [6]\mu_{ad}^{ij}\mu_{be}^{ij}\mu_{df}^{ij} + [9]\mu_{ab}^{ii}\mu_{cd}^{ij}\mu_{ef}^{jj} \end{aligned}$$

Order 6, D=3.

$$\begin{aligned} M_{i^4jk} &= [3]\mu_{ab}^{ii}\mu_{cd}^{ii}\mu_{ef}^{jk} + [12]\mu_{ae}^{ij}\mu_{bf}^{ik}\mu_{cd}^{ii} \\ M_{iiijjk} &= [6]\mu_{ad}^{ij}\mu_{be}^{ij}\mu_{cf}^{ik} + [6]\mu_{ad}^{ij}\mu_{bc}^{ii}\mu_{ef}^{jk} + [3]\mu_{ab}^{ii}\mu_{de}^{jj}\mu_{cf}^{ik} \\ M_{iijjkk} &= \mu_{ab}^{ii}\mu_{cd}^{jj}\mu_{ef}^{kk} + [2]\mu_{ab}^{ii}\mu_{ce}^{jk}\mu_{bf}^{jk} + [2]\mu_{cd}^{jj}\mu_{ae}^{ik}\mu_{bf}^{ik} + [2]\mu_{ef}^{kk}\mu_{ac}^{ij}\mu_{bd}^{ij} \\ &\quad +[8]\mu_{ac}^{ij}\mu_{de}^{jk}\mu_{bf}^{ik} \end{aligned}$$

Order 8, D=2.

$$
\begin{aligned}
M_{i^7j} &= [105]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ij}_{gh} \\
M_{i^6jj} &= [90]\mu^{ij}_{ag}\mu^{ij}_{bh}\mu^{ii}_{cd}\mu^{ii}_{ef} + [15]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jj}_{gh} \\
M_{i^5jjj} &= [60]\mu^{ij}_{af}\mu^{ij}_{bg}\mu^{ij}_{ch}\mu^{ii}_{de} + [45]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jj}_{gh} \\
M_{i^4j^4} &= [9]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{ef}\mu^{jj}_{gh} + [72]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jj}_{gh} + [24]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ij}_{dh}
\end{aligned}
$$

Order 8, D=3.

$$
\begin{aligned}
M_{i^6jk} &= [15]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jk}_{gh} + [90]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eg}\mu^{ik}_{fg} \\
M_{i^5jjk} &= [30]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jk}_{gh} + [60]\mu^{ij}_{af}\mu^{ii}_{bc}\mu^{ik}_{dh} + [15]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{fg}\mu^{ik}_{eh} \\
M_{i^4jjjk} &= [9]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{ef}\mu^{jk}_{gh} + [36]\mu^{jj}_{ef}\mu^{ij}_{ag}\mu^{ik}_{bh}\mu^{ii}_{cd} \\
&\quad +[24]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ik}_{dh} + [36]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jk}_{gh} \\
M_{i^4jjkk} &= [3]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{ef}\mu^{kk}_{gh} + [6]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jk}_{eg}\mu^{jk}_{fh} + [12]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{kk}_{gh} \\
&\quad +[24]\mu^{ik}_{ag}\mu^{ik}_{bg}\mu^{ij}_{ce}\mu^{ij}_{df} + [48]\mu^{ij}_{ae}\mu^{jk}_{fg}\mu^{ik}_{bh}\mu^{ii}_{cd} \\
&\quad +[12]\mu^{ii}_{ab}\mu^{ik}_{bg}\mu^{ik}_{ch}\mu^{jj}_{ef} \\
M_{iiijjkk} &= [9]\mu^{ii}_{ab}\mu^{ij}_{cd}\mu^{jj}_{ef}\mu^{kk}_{gh} + [18]\mu^{ii}_{ab}\mu^{ij}_{cd}\mu^{jk}_{eg}\mu^{jk}_{fh} + [6]\mu^{ij}_{ad}\mu^{ij}_{be}\mu^{ij}_{cf}\mu^{kk}_{gh} \\
&\quad +[18]\mu^{ik}_{ag}\mu^{ik}_{bh}\mu^{ij}_{cd}\mu^{jj}_{ef} + [36]\mu^{ij}_{ad}\mu^{ij}_{be}\mu^{ik}_{cg}\mu^{jk}_{fh} \\
&\quad +[18]\mu^{ik}_{ag}\mu^{jk}_{dh}\mu^{ii}_{bc}\mu^{jj}_{ef}
\end{aligned}
$$

Order 10, D=2.

$$
\begin{aligned}
M_{i^9j} &= [945]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ii}_{gh}\mu^{ij}_{m\ell} \\
M_{i^8jj} &= [105]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ii}_{gh}\mu^{jj}_{m\ell} + [840]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ij}_{gm}\mu^{ij}_{h\ell} \\
M_{i^7jjj} &= [315]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ij}_{gh}\mu^{jj}_{m\ell} + [630]\mu^{ij}_{ah}\mu^{ij}_{bm}\mu^{ij}_{c\ell}\mu^{ii}_{de}\mu^{ii}_{fg} \\
M_{i^6j^4} &= [45]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jj}_{gh}\mu^{jj}_{m\ell} + [360]\mu^{ij}_{ag}\mu^{ij}_{bh}\mu^{ij}_{cm}\mu^{ij}_{d\ell}\mu^{ii}_{ef} \\
&\quad +[540]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eg}\mu^{ij}_{fh}\mu^{jj}_{ml} \\
M_{i^5j^5} &= [120]\mu^{ij}_{af}\mu^{ij}_{bg}\mu^{ij}_{ch}\mu^{ij}_{dm}\mu^{ij}_{e\ell} \\
&\quad +[225]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jj}_{gh}\mu^{jj}_{m\ell} + [600]\mu^{jj}_{fg}\mu^{ij}_{ah}\mu^{ij}_{bm}\mu^{ij}_{c\ell}\mu^{ii}_{de}
\end{aligned}
$$

Order 10, D=3.

$$
\begin{aligned}
M_{i^8jk} &= [105]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ii}_{gh}\mu^{jk}_{m\ell} + [840]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ij}_{gm}\mu^{ik}_{h\ell}\\
M_{i^7jjk} &= [210]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{ij}_{gh}\mu^{jk}_{m\ell} + [630]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eh}\mu^{ij}_{fm}\mu^{ik}_{h\ell}\\
&\quad +[105]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jj}_{hm}\mu^{ik}_{g\ell}\\
M_{i^6jjjk} &= [45]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jj}_{gh}\mu^{jk}_{m\ell} + [270]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eg}\mu^{ij}_{fh}\mu^{jk}_{m\ell}\\
&\quad +[360]\mu^{ij}_{ag}\mu^{ij}_{bh}\mu^{ij}_{cm}\mu^{ik}_{d\ell}\mu^{ii}_{ef} + [270]\mu^{ik}_{a\ell}\mu^{jj}_{gh}\mu^{ij}_{b\ell}\mu^{ii}_{cd}\mu^{ii}_{ef}\\
M_{i^6jjkk} &= [15]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu_{gh}jj\mu^{kk}_{m\ell} + [30]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ii}_{ef}\mu^{jk}_{gm}\mu^{jk}_{h\ell}\\
&\quad +[90]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eg}\mu^{ij}_{fh}\mu^{kk}_{m\ell} + [90]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ik}_{em}\mu^{ik}_{f\ell}\mu^{jj}_{gh}\\
&\quad +[360]\mu^{ii}_{ab}\mu^{ij}_{cg}\mu^{ij}_{dh}\mu^{ik}_{em}\mu^{ik}_{f\ell} + [360]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{eg}\mu^{ik}_{fm}\mu^{jk}_{h\ell}\\
M_{i^5j^4k} &= [45]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{fg}\mu^{jj}_{hm}\mu^{ik}_{e\ell} + [360]\mu^{ii}_{ab}\mu^{ij}_{cf}\mu_{dg}\mu^{jj}_{hm}\mu^{ik}_{e\ell}\\
&\quad +[120]\mu^{ij}_{ag}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ij}_{dh}\mu^{ik}_{e\ell} + [180]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jj}_{gh}\mu^{jk}_{m\ell}\\
&\quad +[240]\mu^{ii}_{ab}\mu^{ij}_{cf}\mu^{ij}_{dg}\mu^{ij}_{eh}\mu^{jk}_{m\ell}\\
M_{i^5jjjkk} &= [45]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jj}_{gh}\mu_{m\ell}kk + [60]\mu^{ii}_{ab}\mu^{ij}_{cf}\mu^{ij}_{dg}\mu^{ij}_{eh}\mu^{kk}_{m\ell}\\
&\quad +[90]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{ij}_{ef}\mu^{jk}_{gm}\mu^{jk}_{h\ell} + 360\mu^{ii}_{ab}\mu^{ij}_{cf}\mu^{ij}_{dg}\mu^{ik}_{em}\mu^{jk}_{h\ell}\\
&\quad +[90]\mu^{ii}_{ab}\mu^{jj}_{fg}\mu^{ik}_{cm}\mu^{jk}_{h\ell} + [180]\mu^{ii}_{ab}\mu^{ij}_{cf}\mu^{jj}_{gh}\mu^{ik}_{dm}\mu^{ik}_{e\ell}\\
&\quad +[120]\mu^{ij}_{af}\mu^{ij}_{bg}\mu^{ij}_{ch}\mu^{ik}_{dm}\mu^{ik}_{e\ell}\\
M_{i^4jjjkkk} &= [27]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{ef}\mu^{jk}_{gh}\mu^{kk}_{m\ell} + [18]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jk}_{eh}\mu^{jk}_{fm}\mu^{ij}_{g\ell}\\
&\quad +[108]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jk}_{gh}\mu^{kk}_{m\ell} + [108]\mu^{ii}_{ab}\mu^{jj}_{ef}\mu^{jk}_{gh}\mu^{ik}_{cm}\mu^{ik}_{d\ell}\\
&\quad +[108]\mu^{ii}_{ab}\mu^{jj}_{ef}\mu^{ij}_{cg}\mu^{ik}_{dh}\mu^{kk}_{m\ell} + [216]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ik}_{dh}\mu^{jk}_{fm}\mu^{jk}_{g\ell}\\
&\quad +[72]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ik}_{dh}\mu^{kk}_{m\ell} + [216]\mu^{ik}_{ah}\mu^{ik}_{bm}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jk}_{g\ell}\\
&\quad +[72]\mu^{ik}_{ah}\mu^{ik}_{bm}\mu^{ik}_{c\ell}\mu^{ij}_{de}\mu^{jj}_{fg}\\
M_{i^4j^4kk} &= [9]\mu^{ii}_{ab}\mu^{ii}_{cd}\mu^{jj}_{ef}\mu^{jj}_{gh}\mu^{kk}_{m\ell} + [72]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jj}_{gh}\mu^{kk}_{m\ell}\\
&\quad +[24]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ij}_{dh}\mu^{kk}_{m\ell} + [36]\mu^{ii}_{ab}\mu^{jj}_{ef}\mu^{jk}_{gm}\mu^{jk}_{h\ell}\\
&\quad +[144]\mu^{ii}_{ab}\mu^{ij}_{ce}\mu^{ij}_{df}\mu^{jk}_{gm}\mu^{jk}_{h\ell} + [36]\mu^{ii}_{ab}\mu^{jj}_{ef}\mu^{jj}_{gh}\mu^{ik}_{cm}\mu^{ik}_{d\ell}\\
&\quad +[144]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{jj}_{gh}\mu^{ik}_{cm}\mu^{ik}_{d\ell} + [288]\mu^{ii}_{ab}\mu^{jj}_{ef}\mu^{ij}_{cg}\mu^{ik}_{dm}\mu^{jk}_{h\ell}\\
&\quad +[192]\mu^{ij}_{ae}\mu^{ij}_{bf}\mu^{ij}_{cg}\mu^{ik}_{dm}\mu^{jk}_{h\ell}
\end{aligned}
$$

A.1.5 PARTICULAR RESULTS WHEN $d = 1$

Here we remind that $\mu^{ij}_{11} = S_{ij}$.

Order 12, d=1, D=2.

$$
\begin{aligned}
M_{i^{11}j} &= 10395S_{ii}^5S_{ij} + 9450S_{ii}^4S_{ij}^2 \\
M_{i^9jjj} &= 2835S_{ii}^4S_{ij}S_{jj} + 7560S_{ii}^3S_{ij}^3 \\
M_{i^8j4} &= 5040S_{ij}^4S_{ii}^2 + 315S_{ii}^4S_{jj}^2 + 5040S_{ii}^3S_{ij}^2S_{jj} \\
M_{i^7j^5} &= 1575S_{ii}^3S_{ij}S_{jj}^2 + 6300S_{ii}^2S_{ij}^3S_{jj} + 2520S_{ii}S_{ij}^5 \\
M_{i^6j^6} &= 720S_{ij}^6 + 225S_{ii}^3S_{jj}^3 + 5400S_{ii}S_{ij}^4S_{jj} + 4050S_{ii}^2S_{ij}^2S_{jj}^2
\end{aligned}
$$

Order 12, d=1, D=3.

$$
\begin{aligned}
M_{i^{10}jk} &= 945S_{ii}^5S_{jk} + 9450S_{ik}S_{ij}S_{ii}^4 \\
M_{ijjk} &= 945S_{ii}^4S_{jj}S_{ik} + 7560S_{ii}^3S_{ij}^2S_{ik} + 1890S_{ii}^4S_{ij}S_{jk} \\
M_{i^8jjjk} &= 315S_{ii}^4S_{jj}S_{jk} + 2520S_{ii}^3S_{ij}S_{jj}S_{ik} + 2520S_{ii}^3S_{ij}^2S_{jk} \\
&\quad +5040S_{ii}^2S_{ij}^3S_{ik} \\
M_{i^7j^4k} &= 315S_{ii}^3S_{jj}^2S_{ik} + 3780S_{ii}^2S_{ij}^2S_{ik} + 1260S_{ii}S_{ij}^4S_{ik} + 1260S_{ii}^3S_{ij}S_{jj}S_{jk} \\
&\quad +3780S_{ii}^2S_{ij}^3S_{jk} \\
M_{i^8jjkk} &= 105S_{ii}^4S_{jj}S_{kk} + 210S_{ii}^4S_{jk}^2 + 840S_{ii}^3S_{ij}^2S_{kk} + 840S_{ii}^3S_{ik}^2S_{jj} \\
&\quad +5040S_{ii}^2S_{ij}^2S_{ik}^2 + 3360S_{ii}^3S_{ik}S_{ij}S_{jk}
\end{aligned}
$$

Order 12, d=1, D=4.

$$
\begin{aligned}
M_{i^4j^4kk\ell\ell} &= 3S_{ii}^2[3S_{jj}^2S_{kk}S_{\ell\ell} + 6S_{jj}^2S_{k\ell}^2 + 12S_{jj}S_{jk}^2S_{\ell\ell} + 24S_{j\ell}^2S_{jk}^2 \\
&\quad +48S_{jk}S_{k\ell}S_{j\ell}S_{jj} + 12S_{jj}S_{j\ell}^2S_{kk}] + 3S_{jj}^2[12S_{ii}S_{ik}^2S_{\ell\ell} \\
&\quad +24S_{il}^2S_{ik}^2 + 48S_{ik}S_{k\ell}S_{i\ell}S_{ii} + 12S_{ii}S_{i\ell}^2S_{kk}] \\
&\quad +24S_{ij}^4S_{kk}S_{\ell\ell} + 48S_{ij}^4S_{k\ell}^2 + 96S_{ij}^3[2S_{ik}S_{jk}S_{\ell\ell} + 2S_{i\ell}S_{j\ell}S_{kk} \\
&\quad +4S_{ik}S_{j\ell}S_{k\ell} + 4S_{i\ell}S_{jk}S_{\ell k}] + 72S_{ij}^2[4S_{ik}^2S_{j\ell}^2 + 4S_{jk}^2S_{i\ell}^2 \\
&\quad +16S_{ik}S_{i\ell}S_{jk}S_{j\ell} + S_{ii}S_{jj}S_{kk}S_{\ell\ell} + 2S_{ii}S_{jj}S_{k\ell}^2] + 12S_{ik}^2[12S_{ji}^2 \\
&\quad \times S_{jj}S_{\ell\ell} + 48S_{ij}S_{i\ell}S_{j\ell}S_{jj} + 12S_{jj}S_{j\ell}^2S_{ii} + 12S_{i\ell}^2[12S_{ji}^2S_{jj}S_{kk} \\
&\quad +48S_{ij}S_{ik}S_{jk}S_{jj} + 12S_{jj}S_{jk}^2S_{ii}] + 12S_{j\ell}^2[12S_{ij}^2S_{ii}S_{kk} \\
&\quad +48S_{ij}S_{jk}S_{ik}S_{ii}] + 12S_{jk}^2[12S_{ij}^2S_{ii}S_{\ell\ell} + 48S_{ij}S_{j\ell}S_{i\ell}S_{ii}] \\
&\quad +576S_{ii}[S_{ik}S_{i\ell}S_{jj}S_{jk}S_{j\ell} + S_{ik}S_{ij}S_{jj}S_{j\ell}S_{k\ell} \\
&\quad +S_{ik}S_{ij}S_{jj}S_{jk}S_{\ell\ell} + S_{il}S_{ij}S_{jj}S_{jk}S_{\ell k} \\
&\quad +S_{i\ell}S_{ij}S_{jj}S_{j\ell}S_{kk} + S_{ik}S_{ij}S_{jj}S_{\ell k}S_{j\ell}]
\end{aligned}
$$

Following the same pattern as the mean, but with more moments involved, the computation of the variance is also conducted.

BIBLIOGRAPHY

[1] Pierre Olivier AMBLARD. "Statistiques d'ordre supérieur pour les signaux non Gaussiens, non linéaires, non stationnaires." PhD thesis. July 1992.

[2] Rafail V Abramov. "The multidimensional maximum entropy moment problem: A review of numerical methods." In: *Communications in Mathematical Sciences* 8.2 (2010), pp. 377–392.

[3] Rex Allen. "Automatic phase pickers: Their present use and future prospects." In: *Bulletin of the Seismological Society of America* 72.6B (1982), S225–S242.

[4] Hassan Amoud, Paul Honeine, Cédric Richard, Pierre Borgnat, and Patrick Flandrin. "Time-frequency learning machines for nonstationarity detection using surrogates." In: *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*. IEEE. 2009, pp. 565–568.

[5] Ulrich Anders and Olaf Korn. "Model selection in neural networks." In: *Neural networks* 12.2 (1999), pp. 309–323.

[6] Theodore W Anderson and Donald A Darling. "Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes." In: *The annals of mathematical statistics* (1952), pp. 193–212. URL: JSTOR.

[7] Nachman Aronszajn. "Theory of reproducing kernels." In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[8] M Baer and U Kradolfer. "An automatic phase picker for local and teleseismic events." In: *Bulletin of the Seismological Society of America* 77.4 (1987), pp. 1437–1445.

[9] Christian Baillard, Wayne Crawford, Valérie Ballu, Clément Hilbert, and Anne Mangeney. "Kurtosis–based P and S phase picker designed for local and regional seismic networks." In: *Bulletin of the Seismological Society of America* 104.1 (2014), pp. 394–409. DOI: 10.1785/0120120347. URL: https://hal.archives-ouvertes.fr/hal-01257944.

[10] M. Basseville. "Detecting changes in signals and systems—a survey." In: *Automatica* 24.3 (1988), pp. 309–326.

[11] M. Basseville and I. Nikiforov. *Detection of Abrupt Changes, Theory and Application*. Information and System Sciences Series. Englewood Cliffs: Prentice-Hall, 1993.

[12] Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*. Vol. 104. prentice Hall Englewood Cliffs, 1993.

[13] Eric Beaucé, William B Frank, Anne Paul, Michel Campillo, and Robert D van der Hilst. "Systematic detection of clustered seismicity beneath the Southwestern Alps." In: *Journal of Geophysical Research: Solid Earth* 124.11 (2019), pp. 11531–11548.

[14] Y. Benjamini and Y. Hochberg. "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing." In: *J. Royal Statist. Soc., Series B* 57 (Nov. 1995), pp. 289–300.

[15] Arthur Berg, Efstathios Paparoditis, and Dimitris N Politis. "A bootstrap test for time series linearity." In: *Journal of Statistical Planning and Inference* 140.12 (2010), pp. 3841–3857.

[16] G D Beskardes, J A Hole, K Wang, M Michaelides, Q Wu, M C Chapman, K K Davenport, L D Brown, and D A Quiros. "A comparison of earthquake backprojection imaging methods for dense local arrays." In: *Geophysical Journal International* 212.3 (Dec. 2017), pp. 1986–2002. ISSN: 0956-540X. DOI: 10.1093/gji/ggx520. eprint: https://academic.oup.com/gji/article-pdf/212/3/1986/23789360/ggx520.pdf. URL: https://doi.org/10.1093/gji/ggx520.

[17] P. Billingsley. *Probability and Measure, 3rd Ed.* Wiley-Interscience, 1995.

[18] Yngve Birkelund, Jarle A Johansen, and Alfred Hanssen. "High-precision surrogate data based tests for gaussianity and linearity of discrete time random processes." In: *2004 12th European Signal Processing Conference*. IEEE. 2004, pp. 73–76.

[19] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. "Learnability and the Vapnik-Chervonenkis dimension." In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965.

[20] Pierre Borgnat, Patrick Flandrin, Paul Honeine, Cédric Richard, and Jun Xiao. "Testing stationarity with surrogates: A time-frequency approach." In: *IEEE Transactions on Signal Processing* 58.7 (2010), pp. 3459–3470.

[21] K. O. Bowman and L. R. Shenton. "Omnibus contours for departures from normality based on b1 and b2." In: *Biometrika* 62 (1975), pp. 243–250.

[22] George Box. "Box and Jenkins: time series analysis, forecasting and control." In: *A Very British Affair*. Springer, 2013, pp. 161–215.

[23] D. R. Brillinger. *Time Series, Data Analysis and Theory*. Holden-Day, 1981.

[24] David R Brillinger. *Time series: data analysis and theory*. SIAM, 2001.

[25] David R Brillinger and Murray Rosenblatt. "Asymptotic theory of estimates of kth-order spectra." In: *Proceedings of the National Academy of Sciences* 57.2 (1967), pp. 206–210.

[26] Judith C Brown. "Calculation of a constant Q spectral transform." In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 425–434.

[27] Joan Bruna and Stéphane Mallat. "Invariant scattering convolution networks." In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886.

[28] YS Cho, Sung Bae Kim, and Edward J Powers. "Time-frequency analysis using AR models with variable forgetting factors." In: *ICASSP*. 1990, pp. 2479–2482.

[29] A. Cichocki and S-I. Amari. *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[30] WB Collis, PR White, and JK Hammond. "Higher-order spectra: the bispectrum and trispectrum." In: *Mechanical systems and signal processing* 12.3 (1998), pp. 375–394.

[31] P. Comon and L. Deruaz. "Normality tests for coloured samples." In: *IEEE-ATHOS Workshop on Higher-Order Statistics*. Begur, Spain, June 1995, pp. 217–221.

[32] P. Comon and C. Jutten, eds. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Oxford, Burlington: Academic Press, 2010.

[33] Pierre Comon. "Quelques développements récents en traitement du signal." Habilitation à diriger des recherches. Université Nice Sophia Antipolis, Sept. 1995. URL: https://theses.hal.science/tel-00473197.

[34] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[35] Harald Cramér. "A contribution to the theory of statistical estimation." In: *Scandinavian Actuarial Journal* 1946.1 (1946), pp. 85–94. DOI: 10.1080/03461238.1946.10419631.

[36] R. D'agostino. "An Omnibus test of Normality for moderate and large size samples." In: *Biometrika* 58.2 (1971), pp. 341–348.

[37] R. D'agostino and E. S. Pearson. "Tests for departure from Normality. Empirical results for the distribution of b2 and b1." In: *Biometrika* 60.3 (1973), pp. 613–622.

[38] K Drouiche. "A new test for whiteness." In: *IEEE transactions on signal processing* 48.7 (2000), pp. 1864–1871.

[39] James Durbin. "The fitting of time-series models." In: *Revue de l'Institut International de Statistique* (1960), pp. 233–244.

[40] Paul S Dysart and Jay J Pulli. "Regional seismic event classification at the NORESS array: seismological measurements and the use of trained neural networks." In: *Bulletin of the Seismological Society of America* 80.6B (1990), pp. 1910–1933.

[41] George S Easton and Robert E McCulloch. "A multivariate generalization of quantile-quantile plots." In: *Journal of the American Statistical Association* 85.410 (1990), pp. 376–386.

[42] Bruno Ebner and Norbert Henze. "Tests for multivariate normality—a critical review with emphasis on weighted L2 L 2-statistics." In: *Test* 29.4 (2020), pp. 845–892.

[43] Sara El Bouch, Olivier Michel, and Pierre Comon. "A normality test for multivariate dependent samples." In: *Signal Processing* 201 (2022), p. 108705.

[44] S. ElBouch, O. Michel, and P. Comon. "Joint Normality Test Via Two-Dimensional Projection." In: *ICASSP*. hal-03369151. Singapore, 2022.

[45] Sara Elbouch. "Supplementary material." working paper or preprint. Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03343508.

[46] Sara Elbouch, Olivier Michel, and Pierre Comon. "Multivariate Normality test for colored data." In: *European Signal Processing Conference*. Belgrade, Serbia, Aug. 2022. URL: https://hal.archives-ouvertes.fr/hal-03691615.

[47] Sara Elbouch, Olivier Michel, and Pierre Comon. "Un Test de Normalité pour les Processus Colorés Multivariés." In: *GRETSI*. Nancy, 2022.

[48] T. W. Epps. "Testing that a stationary time series is Gaussian." In: *The Annals of Statistics* 15.4 (1987), pp. 1683–1698.

[49] Durdu Ömer Faruk. "A hybrid neural network and ARIMA model for water quality time series prediction." In: *Engineering applications of artificial intelligence* 23.4 (2010), pp. 586–594.

[50] Nicholas I Fisher. "Graphical methods in nonparametric statistics: A review and annotated bibliography." In: *International Statistical Review/Revue Internationale de Statistique* (1983), pp. 25–58.

[51] Patrick Flandrin. *Time-frequency/time-scale analysis*. Academic press, 1998.

[52] W. B. Frank and N. M. Shapiro. "Automatic detection of low-frequency earthquakes (LFEs) based on a beamformed network response." In: *Geophysical Journal International* 197.2 (Mar. 2014), pp. 1215–1223. ISSN: 0956-540X. DOI: 10.1093/gji/ggu058. eprint: https://academic.oup.com/gji/article-pdf/197/2/1215/17366527/ggu058.pdf. URL: https://doi.org/10.1093/gji/ggu058.

[53] David Freedman and Persi Diaconis. "On the histogram as a density estimator: L 2 theory." In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57.4 (1981), pp. 453–476.

[54] T. Gasser. "Goodness-of-fit tests for correlated data." In: *Biometrika* 62.3 (1975), pp. 563–570.

[55] Georgios B Giannakis and Erchin Serpedin. "A bibliography on nonlinear system identification." In: *Signal Processing* 81.3 (2001), pp. 533–580.

[56] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[57] Ciprian Doru Giurcăneanu and Seyed Alireza Razavi. "AR order selection in the case when the model parameters are estimated by forgetting factor least-squares algorithms." In: *Signal Processing* 90.2 (2010), pp. 451–466.

[58] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016.

[59] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. "Knowledge distillation: A survey." In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819.

[60] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. "LSTM: A search space odyssey." In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.

[61] G Gutenberg and CF Richter. "Seismicity of the earth and associated phenomena, Howard Tatel." In: *Journal of Geophysical Research* 55 (1950), p. 97.

[62] E. J. (Edward James) Hannan. 1970.

[63] David I Harvey and Stephen J Leybourne. "Testing for time series linearity." In: *The Econometrics Journal* 10.1 (2007), pp. 149–165.

[64] Nima Hatami, Yann Gavet, and Johan Debayle. "Classification of time-series images using deep convolutional neural networks." In: *Tenth international conference on machine vision (ICMV 2017)*. Vol. 10696. SPIE. 2018, pp. 242–249.

[65] David Haussler. "Decision theoretic generalizations of the PAC model for neural net and other learning applications." In: *The Mathematics of Generalization*. CRC Press, 2018, pp. 37–116.

[66] S. Haykin. *Unsupervised Adaptive Filtering*. Vol. 1 & 2. series in Adaptive and Learning Systems for Communications, Signal Processing, and Control. Wiley, 2000.

[67] R. Henze. "Invariant tests for multivariate normality: a critical review." In: *Statistical papers* 43 (2002), pp. 467–506.

[68] M. Hinich. "Testing for Gaussianity and Linearity of a Stationary Time Series." In: *Jour. Time Series Analysis* 3.3 (1982), pp. 169–176.

[69] David V Hinkley. "Inference about the change-point from cumulative sum tests." In: *Biometrika* 58.3 (1971), pp. 509–523.

[70] Marius Hofert. "Sampling archimedean copulas." In: *Computational Statistics & Data Analysis* 52.12 (2008), pp. 5163–5174.

[71] R. V. Hogg. "Statistical robustness: One view of its use in applications today." In: *The American Statistician* 33.3 (1979), pp. 108–115.

[72] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." In: *Neural networks* 4.2 (1991), pp. 251–257.

[73] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. "Deep learning for time series classification: a review." In: *Data mining and knowledge discovery* 33.4 (2019), pp. 917–963.

[74] Herbert Jaeger and Harald Haas. "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication." In: *science* 304.5667 (2004), pp. 78–80.

[75] Carlos M Jarque and Anil K Bera. "A test for normality of observations and regression residuals." In: *Int. Statistical Review* (1987), pp. 163–172.

[76] Edwin T Jaynes. "On the rationale of maximum-entropy methods." In: *Proceedings of the IEEE* 70.9 (1982), pp. 939–952.

[77] Maya Kallas, Paul Honeine, Clovis Francis, and Hassan Amoud. "Kernel autoregressive models using Yule–Walker equations." In: *Signal Processing* 93.11 (2013), pp. 3053–3061.

[78] Maya Kallas, Paul Honeine, Cédric Richard, Clovis Francis, and Hassan Amoud. "Prediction of time series using yule-walker equations with kernels." In: *Proc. 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp. 2185 –2188. DOI: 10.1109/ICASSP.2012.6288346. URL: https://hal.archives-ouvertes.fr/hal-01966015.

[79] H Karabulut, Jean Schmittbuhl, S Özalaybey, Olivier Lengline, A Kömeç-Mutlu, Virginie Durand, Michel Bouchon, G Daniel, and MP Bouin. "Evolution of the seismicity in the eastern Marmara Sea a decade before and after the 17 August 1999 Izmit earthquake." In: *Tectonophysics* 510.1-2 (2011), pp. 17–27.

[80] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1997.

[81] Maurice George Kendall, Alan Stuart, and J Keith Ord. *Kendall's advanced theory of statistics*. Oxford University Press, Inc., 1987.

[82] Mehdi Khashei and Mehdi Bijari. "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting." In: *Applied soft computing* 11.2 (2011), pp. 2664–2675.

[83] Tõnu Kollo. "Multivariate skewness and kurtosis measures with an application in ICA." In: *Journal of Multivariate Analysis* 99.10 (2008), pp. 2328–2338.

[84] Andrei Nikolaevich Kolmogorov and Yu A Rozanov. "On strong mixing conditions for stationary Gaussian processes." In: *Theory of Probability & Its Applications* 5.2 (1960), pp. 204–208.

[85] Qingkai Kong, Daniel Trugman, Zachary Ross, Michael Bianco, Brendan Meade, and Peter Gerstoft. "Machine Learning in Seismology: Turning Data into Insights." In: *Seismological Research Letters* 90 (Nov. 2018). DOI: 10.1785/0220180259.

[86] James A Koziol. "A note on the asymptotic distribution of Mardia's measure of multivariate kurtosis." In: *Communications in Statistics-Theory and Methods* 15.5 (1986), pp. 1507–1513.

[87] Jean-Louis Lacoume, Pierre-Olivier Amblard, and Pierre Comon. *Statistiques d'ordre supérieur pour le traitement du signal*. Masson, 1997.

[88] Nadège Langet, Alessia Maggi, Alberto Michelini, and Florent Brenguier. "Continuous Kurtosis-Based Migration for Seismic Event Detection and Location, with Application to Piton de la Fournaise Volcano, La Réunion." In: *Bulletin of the Seismological Society of America* 104.1 (2014), pp. 229–246.

[89] M Leonard and BLN Kennett. "Multi-component autoregressive techniques for the analysis of seismograms." In: *Physics of the Earth and Planetary Interiors* 113.1-4 (1999), pp. 247–263.

[90] Jiajuan Liang, Runze Li, Hongbin Fang, and Kai-Tai Fang. "Testing multinormality based on low-dimensional projection." In: *Journal of Statistical Planning and Inference* 86.1 (2000), pp. 129–141.

[91] Zhishuai Liu, Guihua Yao, Qing Zhang, Junpu Zhang, and Xueying Zeng. "Wavelet scattering transform for ECG beat classification." In: *Computational and Mathematical Methods in Medicine* 2020 (2020).

[92] I. N. Lobato and C. Velasco. "A Simple Test of Normality for Time Series." In: *Econometric Theory* 20.4 (2004), pp. 671–689. URL: http://www.jstor.org/stable/3533541.

[93] H. Lütkepohl. *Introduction to multiple time series analysis*. Springer Science & Business Media, 2013.

[94] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Berlin [u.a.]: Springer, 2005. XXI, 764. ISBN: 3540262393. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+366296310&sourceid=fbw_bibsonomy.

[95] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal, et al. "Long short term memory networks for anomaly detection in time series." In: *Proceedings*. Vol. 89. 2015, pp. 89–94.

[96] James Francis Malkovich and A Afifi. "On tests for multivariate normality." In: *Journal of the American statistical association* 68.341 (1973), pp. 176–179.

[97] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[98] K. V. Mardia. "Measures of Multivariate skewness and kurtosis with applications." In: *Biometrika* 57 (1970), pp. 519–530.

[99] K. V. Mardia. "Tests of Univariate and Multivariate Normality." In: *Handbook of Statistics, Vol.1*. Ed. by P. R. Krishnaiah. North-Holland, 1980, pp. 279–320.

[100] Kanti V Mardia and K Foster. "Omnibus tests of multinormality based on skewness and kurtosis." In: *Communications in Statistics-theory and methods* 12.2 (1983), pp. 207–221.

[101] P. Mccullagh. *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.

[102] Christopher J Mecklin and Daniel J Mundfrom. "An appraisal and bibliography of tests for multivariate normality." In: *International Statistical Review* 72.1 (2004), pp. 123–138.

[103] D. S. Moore. "A Chi-Square Statistic with Random cell boundaries." In: *The Annals of Statistics* 42.1 (1971), pp. 147–156.

[104] D. S. Moore. "The effect of dependence on Chi squared tests of fit." In: *The Annals of Statistics* 10.4 (1982), pp. 1163–1171.

[105] Tamás F Móri, Vijay K Rohatgi, and GJ Székely. "On multivariate skewness and kurtosis." In: *Theory of Probability & Its Applications* 38.3 (1994), pp. 547–551.

[106] E. Moulines, K. Choukri, and M. Charbit. "Testing that a multivariate stationary time series is Gaussian." In: *Sixth SSAP Workshop on Stat. Signal and Array Proc.* 1992, pp. 185–188.

[107] S Mostafa Mousavi, William L Ellsworth, Weiqiang Zhu, Lindsay Y Chuang, and Gregory C Beroza. "Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking." In: *Nature communications* 11.1 (2020), pp. 1–12.

[108] S Mostafa Mousavi, Weiqiang Zhu, William Ellsworth, and Gregory Beroza. "Unsupervised clustering of seismic signals using deep convolutional autoencoders." In: *IEEE Geoscience and Remote Sensing Letters* 16.11 (2019), pp. 1693–1697.

[109] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. "Network information criterion-determining the number of hidden units for an artificial neural network model." In: *IEEE transactions on neural networks* 5.6 (1994), pp. 865–872.

[110] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[111] A. Nieto-Reyes, J. A. Cuesta-Albertos, and F. Gamboa. "A random-projection based test of Gaussianity for stationary processes." In: *Computational Statistics & Data Analysis* 75 (2014), pp. 124–141.

[112] C. L. Nikias and A. P. Petropulu. *Higher-Order Spectra Analysis*. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1993.

[113] Chrysostomos L Nikias and Jerry M Mendel. "Signal processing with higher-order spectra." In: *IEEE Signal processing magazine* 10.3 (1993), pp. 10–37.

[114] Chester H Page. "Instantaneous power spectra." In: *Journal of Applied Physics* 23.1 (1952), pp. 103–106.

[115] Ewan S Page. "Continuous inspection schemes." In: *Biometrika* 41.1/2 (1954), pp. 100–115.

[116] Piero Poli. "Creep and slip: seismic precursors to the Nuugaatsiaq landslide (Greenland): seismic precursors to a landslide." In: *Geophysical Research Letters* (Aug. 2017). DOI: 10.1002/2017GL075039.

[117] H Vincent Poor and Olympia Hadjiliadis. *Quickest detection*. Cambridge University Press, 2008.

[118] Zacharias Psaradakis and Marián Vávra. "A distance test of normality for a wide class of stationary processes." In: *Econometrics and Statistics* 2 (2017), pp. 50–60.

[119] R. Reed. "Pruning algorithms-a survey." In: *IEEE Transactions on Neural Networks* 4.5 (1993), pp. 740–747. DOI: 10.1109/72.248452.

[120] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. Vol. 15. Series in Computer Science. London: World Scientific Publ., 1989.

[121] Murray Rosenblatt. "A central limit theorem and a strong mixing condition." In: *Proceedings of the national Academy of Sciences* 42.1 (1956), pp. 43–47.

[122] Bertrand Rouet-Leduc, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J Humphreys, and Paul A Johnson. "Machine learning predicts laboratory earthquakes." In: *Geophysical Research Letters* 44.18 (2017), pp. 9276–9282.

[123] Baptiste Rousset, Roland Burgmann, and Michel Campillo. "Slow slip events in the roots of the San Andreas fault." In: *Science Advances* 5 (Feb. 2019). DOI: 10.1126/sciadv.aav3274.

[124] Elena Rusticelli, Richard A Ashley, Estela Bee Dagum, and Douglas M Patterson. "A new bispectral test for nonlinear serial dependence." In: *Econometric Reviews* 28.1-3 (2008), pp. 279–293.

[125] Christos D Saragiotis, Leontios J Hadjileontiadis, and Stavros M Panas. "PAI-S/K: A robust automatic seismic P phase arrival identification scheme." In: *IEEE Transactions on Geoscience and Remote Sensing* 40.6 (2002), pp. 1395–1404.

[126] Brent Scarff. *Atomic force microscopy*. 2013. URL: https://towardsdatascience.com/understanding-backpropagation-abcc509ca9d0.

[127] Martin Schetzen. "Nonlinear system modeling based on the Wiener theory." In: *Proceedings of the IEEE* 69.12 (1981), pp. 1557–1573.

[128] Raghavendra Selvan. "Bayesian tracking of multiple point targets using Expectation Maximization." PhD thesis. July 2015.

[129] Léonard Seydoux, Randall Balestriero, Piero Poli, Maarten de Hoop, Michel Campillo, and Richard Baraniuk. "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning." In: *Nature communications* 11.1 (2020), pp. 1–12.

[130] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[131] Samuel S Shapiro and RS Francia. "An approximate analysis of variance test for normality." In: *Journal of the American statistical Association* 67.337 (1972), pp. 215–216.

[132] Laurent Sifre and Stéphane Mallat. "Rigid-motion scattering for texture classification." In: *arXiv preprint arXiv:1403.1687* (2014).

[133] Rene Steinmann, Leonard Seydoux, Eric Beaucé, and Michel Campillo. "Hierarchical exploration of continuous seismograms with unsupervised learning." In: *Journal of Geophysical Research: Solid Earth* 127.1 (2022), e2021JB022455.

[134] Petre Stoica. "A test for whiteness." In: *IEEE transactions on automatic control* 22.6 (1977), pp. 992–993.

[135] Mervyn Stone. "Cross-validation: A review." In: *Statistics: A Journal of Theoretical and Applied Statistics* 9.1 (1978), pp. 127–139.

[136] Tetsuo Takanami and Genshiro Kitagawa. "Multivariate time-series model to estimate the arrival times of S-waves." In: *Computers & Geosciences* 19.2 (1993), pp. 295–301.

[137] Gerald Tesauro. "Practical issues in temporal difference learning." In: *Advances in neural information processing systems* 4 (1991).

[138] Leslie G Valiant. "A theory of the learnable." In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[139] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I.* Wiley, 1968, 2001.

[140] Peter Whittle. "On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix." In: *Biometrika* 50.1-2 (1963), pp. 129–134.

[141] Woon Wong. "Frequency domain tests of multivariate Gaussianity and linearity." In: *Journal of time series analysis* 18.2 (1997), pp. 181–194.

[142] Takayuki Yamada and Tetsuto Himeno. "Estimation of multivariate 3rd moment for high-dimensional data and its application for testing multivariate normality." In: *Computational Statistics* 34.2 (2019), pp. 911–941.

[143] Izzet B Yildiz, Herbert Jaeger, and Stefan J Kiebel. "Re-visiting the echo state property." In: *Neural networks* 35 (2012), pp. 1–9.

[144] Arnold Zellner. "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." In: *Journal of the American statistical Association* 57.298 (1962), pp. 348–368.

[145] Weiqiang Zhu and Gregory C Beroza. "PhaseNet: a deep-neural-network-based seismic arrival-time picking method." In: *Geophysical Journal International* 216.1 (Oct. 2018), pp. 261–273. ISSN: 0956-540X. DOI: 10.1093/gji/ggy423. eprint: https://academic.oup.com/gji/article-pdf/216/1/261/26329430/ggy423.pdf. URL: https://doi.org/10.1093/gji/ggy423.

[146] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net." In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

DECLARATION

Toutes les phrases, passages et illustrations cités dans ce document provenant de travaux d'autres personnes ont été spécifiquement identifiés ou référencées par une citation claire à l'auteur, le travail et la ou les pages correspondantes. Je comprends que faillir à cela équivaut à du plagiat et sera considéré en tant que tel lors de l'évaluation.

Grenoble, France, October 2022

Sara El Bouch

RÉSUMÉ SUBSTANTIEL

INTRODUCTION

Le travail présenté dans ce manuscrit est ancré dans une discipline: Le traitement du signal à la frontière des mathématiques appliquées, de la physique et de l'informatique. Cette discipline est depuis longtemps sollicitée pour résoudre divers problèmes dans à peu près tous les domaines générant un *signal*; comment extraire des informations d'un signal bruité ? Comment exploiter des connaissances a priori pour rendre l'extraction plus fine ? Les réponses à ces questions ont naturellement introduit des outils à saveur inférentielle en tentant de construire un modèle (paramétrique ou non-paramétrique) régissant les observations. La quantité en forte croissance des données et les progrès réalisés dans le domaine de l'informatique entraînent la prolifération d'une quantité sans précédent de modèles. Leur objectif est de parcourir des tonnes de données et d'en extraire des informations utiles en contournant l'intervention humaine. Cet exercice se déroule sous le couvert du ML. Les pratiques du traitement du signal et du ML ont les mêmes objectifs et sont généralement combinées de manière favorable pour extraire efficacement des informations des données.

Nouveau paradigme proéminent

La question au cœur de ce manuscrit est : *"Comment détecter une série temporelle de faible amplitude noyée dans le bruit ?"*.. La première difficulté du problème à résoudre réside dans la présence de bruit ambient qui réussit à masquer les séries de faible amplitude, et en deuxième lieu, dans l'absence d'informations *a priori* sur la source qui génère les signaux à détecter.

Détection et estimation

Il s'agit essentiellement d'un problème interdisciplinaire auquel des outils adaptés ont été développées dans le domaine du traitement du signal, et récemment de plus en plus en utilisant des pratiques de ML. Il nous a donc semblé naturel de commencer par quelques piliers du traitement du signal et de présenter le contexte dans lequel il a le plus fonctionné, régi par trois propriétés : Linéarité, stationnarité et Gaussianité, puis de passer progressivement aux outils adéquats lorsqu'au moins une de ces propriétés est écartée.

Avant de tenter de répondre à la question ci-dessus, nous ajoutons d'autres contraintes à la solution proposée : Le détecteur doit être efficace, d'un point de vue théorique et computationnel. Il doit être accompagné de garanties sur le taux de fausses alarmes. La charge de

Détecteur à faible charge de calcul

calcul doit être faible pour permettre le traitement de grands ensembles de données.

Satisfaire la première contrainte (garanties théoriques sur le taux de fausses alarmes) nous a conduit dans le domaine des statistiques. Afin de fournir des résultats statistiquement significatifs, nous devons choisir rigoureusement une statistique de test et définir sa distribution asymptotique sous un ensemble d'hypothèses. Étonnamment, il n'existait pas de procédures efficaces en termes de calcul pour tester qu'une série temporelle multivariée est Gaussienne, et notre quête s'est transformée en une contribution.

Taux de fausse alarme fixe

La principale préoccupation était et reste de fournir un détecteur opérationnel, à cette fin, nous traduisons les principaux résultats de nos contributions en un détecteur de changement séquentiel. Nous concluons sur les performances du détecteur sur un ensemble d'expériences numériques. Nous passons progressivement des données générées synthétiquement aux données du monde réel.

En fait, le travail dans ce manuscrit est également ancré dans les applications physiques en sismologie. Nous nous concentrons sur une application qui a une longue et riche histoire dans ce domaine : la détection des secousses sismiques. La loi de Gutenberg-Richter [61] stipule que le nombre cumulé de tremblements de terre augmente exponentiellement avec une magnitude décroissante. Ces événements sont sévèrement noyés par le bourdonnement sismique, les détecter n'est pas une tâche facile mais une tâche gratifiante car elle conduira à dévoiler de nouvelles formes d'ondes sismique, et par extension à une meilleure compréhension de la dynamique des tremblements de terre.

Dans ce cadre, le but étant de détecter avec un minimum d'a priori, on ne peut pas avoir recours à de larges données étiquetées. Nous proposons d'utiliser notre détecteur afin de fournir des événements nouvellement détectés. Cependant, ceci soulève d'importantes questions sur l'évaluation de différentes méthodes dans ce contexte. Comment comparer des modèles sur des instances rares ? Comment comparer différents catalogues subjectivement construits ?

Les différentes facettes de ce travail peuvent peut-être être déroutantes au premier abord. Il s'agira d'un va-et-vient entre les méthodes de traitement du signal, les tests d'hypothèses statistiques et le sondage d'événements avec ML. Ensuite, l'applicabilité pratique de la principale contribution sera mise à l'épreuve sur des données du monde réel. Les articulations de ce manuscrit peuvent être résumées comme suit :

Piliers de l'analyse des séries temporelles (y compris les pratiques de ML) $\Longleftrightarrow$ Test de normalité pour les séries temporelles multivariées $\Longleftrightarrow$ Détecteur opérationnel sur données synthétiques et réelles (exclusivement sur des applications en sismologie).

Pour aider le lecteur, nous détaillons dans ce qui suit le plan et nos principales contributions.

OUTLINE ET CONTRIBUTIONS

Le chapitre 1 introduit les définitions et les théorèmes de l'analyse des séries temporelles afin de fournir au lecteur les outils nécessaires pour mieux comprendre notre cadre et nos contributions. Les problèmes de traitement du signal sont généralement résolus en supposant un modèle probabiliste sous-jacent. Nous développons plus en détail cette saveur inférentielle dans le chapitre suivant 2. La partie **??** sert à poser le cadre de notre travail, et introduit en même temps les éléments constitutifs de nos contributions. Nous étions partagés entre présenter autant d'outils que possible et nous concentrer sur ceux qui réapparaîtront dans la suite. En raison de la prolifération continue de modèles pour différents types de données, nous nous concentrons uniquement sur les méthodes pertinentes pour les séries temporelles, et si possible les séries temporelles multivariées. La partie ii est consacrée à nos contributions originales, dans lesquelles nous proposons une méthode pertinente pour évaluer la gaussianité des séries temporelles multivariées. Nous l'ouvrons par le chapitre 3 qui fournit une vue d'ensemble des tests de normalité préexistants et motive la nécessité d'en dériver un nouveau dans notre cadre. À des fins de calcul, nous nous concentrons sur les statistiques d'ordre supérieur, plus précisément le Kurtosis de Mardia, et nous définissons complètement la distribution limite de cette statistique de test (sous gaussianité) pour les séries temporelles colorées. Le contexte théorique et les étapes du calcul sont nos premiers travaux [43] :

Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *A normality test for multivariate dependent samples* in Signal Processing, Elsevier 2022

dans lequel les outils nécessaires, les théorèmes et les principaux résultats ont été détaillés. Le chapitre 4 reproduit de nombreux paragraphes et équations du [43]. La principale préoccupation du chapitre 5 est maintenant de traduire les résultats théoriques en un détecteur opérationnel en temps réel. Une étude préliminaire a d'abord exam-

iné la généralisation de nos résultats sur les processus bivariés au cas général d-varié avec notre deuxième contribution : [44].

> Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Joint Normality Test Via Two-dimensional projections* in ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, Singapour 2022

dans lequel nous avons mené une étude comparative entre le test de normalité conjoint appliqué sur des projections aléatoires 2D et son homologue scalaire sur une projection unidimensionnelle. Les résultats sont discutés à la fin du chapitre 4. Disposant désormais d'un outil pratique pour évaluer la normalité, nous continuons à tester les performances du test avec ou sans préblanchiment ; nous souhaitons tester la normalité des résidus de régression plutôt que des données brutes. Sur la base des résultats de [46] :, nous proposons en

> Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Multivariate Normality Test for Colored data* in EUSIPCO European Signal processing community, Belgrade, Serbia 2022

un détecteur séquentiel opérationnel en deux étapes. Avant d'appliquer notre test, les données entrantes sont préblanchies à l'aide d'un modèle auto-régressif multidimensionnel. Des expériences numériques ont été menées pour évaluer la puissance du test en tenant compte de la dépendance spatiale et temporelle du processus. Encouragés par les résultats sur les données synthétiques, nous avons poursuivi les simulations sur de petites portions de données du monde réel dans [47] :

> Sara El Bouch, Olivier J.J. Michel, Pierre Comon, *Un Test de Normalité pour les Processus Colorés Multivariés* au GRETSI, Nancy, France 2022

Ces résultats sont présentés dans le chapitre 5. La dernière et troisième partie de ce travail est la partie iii dans laquelle le chapitre 6 traite de la fusion de nos travaux avec les expériences sismologiques. Le domaine est historiquement riche en méthodes de détection, et récemment il a attiré les pratiques ML. Il est donc nécessaire de passer en revue certaines des méthodes pertinentes pour notre tâche de détection. Nous poursuivons ensuite avec la simulation sur des données réelles.

INDEX