

Faculté des sciences

Classification of Spectral Densities of Stationary Time Series

Author: **Mirco LESCART**

Supervisor: **Rainer VON SACHS**

Reader: **Christian HAFNER**

Academic year 2023–2024

Master [120] en science des données, orientation statistique

Acknowledgments.

I would like to express my sincere gratitude to my promotor, Professor Rainer von Sachs, for his unwavering support and guidance throughout the development of this thesis. His expertise, creativity, and openness to new ideas were invaluable in shaping the direction of my research, and I greatly appreciated his availability and dedication.

I also wish to thank Gabriel Bailly for his consistent assistance and support during this process. His help was an essential part of my journey.

I am deeply thankful to my family, friends, and loved ones for their encouragement and support throughout my studies.

Finally, I would like to thank Christian Haffner for his time and effort in reviewing my thesis. I am also sincerely grateful to Eric Hénaut for his thorough proofreading and polishing of my thesis, which greatly contributed to its final form.

Contents

1	Introduction	1
1.1	Overview of Time Series Classification	1
1.2	Contributions and Methodology	2
2	Time Series Theory	4
2.1	Definition and Statistical Properties	6
2.1.1	Definition of Time Series	6
2.1.2	Statistical Properties of Time Series	7
3	Spectral Analysis of Time Series	11
3.1	Spectral Density	12
3.2	Estimator of the Spectral Density	14
3.2.1	Periodogram	14
3.2.2	Smoothed Periodogram	16
4	Geometry of Spectral Density	22
4.1	Theoretical Foundations of Differential Geometry	23
4.1.1	Manifolds	23
4.1.2	Riemannian Geometry	24
4.2	Spectral Density Matrices and Manifolds	25
4.2.1	Karcher Mean	27
5	Method of classification	30
5.1	k-Nearest Neighbors	30
5.2	Minimum Distance to Means	33
5.3	Sensitivity to Outliers	35
6	Spectral Distances	37
6.1	Introduction to Spectral Distances	37
6.2	Mathematical Framework for Spectral Distance Comparison	38
6.3	Euclidean Distance	39
6.4	Affine-Invariant Riemannian (AIR) Distance	40
6.5	Log-Euclidean Distance	42

6.6	Whittle Distance	42
6.6.1	Whittle Consistency	44
6.7	Comparative Analysis of Karcher Means Associated with Different Spectral Distances	50
7	Application : Simulation Study	52
7.1	Vector Autoregression (VAR)	52
7.2	Data Generation	53
7.3	Frequency Domain	55
7.4	Classification	57
7.4.1	Classification Methods and Distance Metrics	57
7.4.2	Results	58
8	Application : Electroencephalography	61
8.1	Data Description	62
8.2	Frequency Domain Analysis	64
8.3	Classification	65
8.3.1	Subject-Specific Classification	68
8.3.2	Cross-Subject Classification	71
8.4	Comparison with Previous Research	75
8.5	Conclusion & Discussion	76
9	Conclusion and Future Work	78
9.1	Conclusion	78
9.2	Discussion and Future Work	78

List of Notations

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

$O(\cdot)$	Big O notation
$(A)_{(i,j)}$	Element (i, j) of the matrix A
$\bar{a}$	Conjugate of the complex number a
$\mathbb{C}^{P \times P}$	$P \times P$ complex matrices set
$\Gamma_{\mathbf{X}}$	Covariance matrix of the time series $\mathbf{X}$
$\mathcal{C}_l$	Class l
$\mathcal{H}_P$	$P \times P$ Hermitian matrices set
$\mathcal{H}_P^+$	$P \times P$ Hermitian Positive Definite Matrices set
Ω	Subset of frequencies
ω	Frequency
ϕ_l	True spectral density of the class $\mathcal{C}_l$
A^*	Conjugate transpose of the complex matrix A
A^T	Transpose of the matrix A
B	Number of Tapering Functions
$d_{AIR}(f_X, f_Y)$	Affine Invariant Riemannian distance between spectral densities f_X, f_Y
$d_E(f_X, f_Y)$	Euclidean distance between spectral densities f_X, f_Y
$d_{LE}(f_X, f_Y)$	Log-Euclidean distance between spectral densities f_X, f_Y
$d_W(f_X, f_Y)$	Whittle distance between spectral densities f_X, f_Y
$f_X(\cdot)$	Spectral density of the time series X

$G_d(\cdot, \dots, \cdot)$	Karcher mean associated with the distance d
$I_X(\cdot)$	Periodogram of the time series X
K	Number of folds in cross-validation (upper-case)
k	Number of neighbors (lower-case)
n_w	Time bandwidth

Chapter 1

Introduction

1.1 Overview of Time Series Classification

Time series data are prevalent across diverse domains such as financial markets, climatic variability [14], medical diagnostics [1, 28], and daily activities. It captures the evolution of phenomena over time, providing insights into past behaviors and future trends. However, classifying time series based on their properties can be challenging. For instance, distinguishing between a healthy heartbeat and an erratic one, or differentiating stable versus volatile stock market trends, highlight the importance of effective time series classification.

Time series classification aims to identify a series as belonging to a specific group using labeled training data. For example, Holan et al. (2010) [16] classify animal behavior from time series data, while Lines et al. (2011) [31] categorize household devices based on electricity usage patterns. These applications show the broad utility of time series classification in various fields, from behavioral analysis to smart technology.

This thesis investigates an advanced approach to time series classification, moving the focus from the conventional time-domain analysis to the frequency domain. While traditional methods often analyze time series based on how data points change over time, this research examines how different frequency components contribute to the overall behavior of the time series. This perspective represents not merely a different viewpoint but a significant shift that reveals new opportunities for classifying complex time series that might appear similar in the time domain but reveal distinctive features in the frequency domain.

The field of time series classification is well-established, encompassing a variety of methods that are tailored to different data types and classification objectives. Approaches may be classified according to whether they analyse raw time series values or transformed data. Such techniques include advanced methods such as deep neural networks [21, 34] and dynamic time warping (DTW) [30], which align time series to accommodate shifts and distortions. This thesis, however, focuses on the spectral densities of time series, which represent the distribution of data variance across various frequencies.

Spectral densities play an instrumental role in uncovering the underlying processes that generate time series data, particularly when the data is discrete, real-valued, weakly

stationary, and has an absolutely summable autocovariance function. These conditions facilitate the application of spectral analysis, which allows for the capture of frequency characteristics through advanced methods such as the multitaper spectral estimator. The objective is to categorize multiple time series based on their estimated spectral densities, with each category corresponding to an underlying process represented by a true spectral density. As the actual spectral densities are often unknown and must be estimated from the data, it is essential to compare these estimates using various distance metrics in order to determine the similarity between time series and thereby inform their classification into predefined categories.

1.2 Contributions and Methodology

This thesis aims to make several key contributions to the field of time series classification by leveraging frequency domain analysis of spectral densities. At the core of this research is an investigation into the geometrical space in which spectral densities reside. By employing Riemannian geometry, this study examines how the curvature and geometric structure of spectral density matrices influence the performance of distance metrics utilized in classification tasks. Spectral densities typically inhabit complex manifolds rather than simple Euclidean spaces, necessitating specific tools for effective analysis. This aspect of the research is thoroughly examined in Chapter 4.

A notable contribution of this research is the comprehensive evaluation and comparison of various distance metrics for spectral densities. It is crucial to use metrics that define the similarity between time series while accounting for the geometrical structure of spectral densities. By applying a Riemannian framework, this research identifies the optimal distances for classification tasks, particularly those that leverage the manifold structure of spectral densities. Chapter 6 presents these findings along with theoretical results demonstrating the consistency of one distance metric in conjunction with a specific classification method, Minimum Distance to Means (MDM).

Furthermore, the robustness of different spectral distance metrics to outliers is empirically examined. Outliers can significantly distort the spectral characteristics of time series, potentially leading to misclassifications. Analyzing how various metrics respond to such anomalies provides insights into which distances offer greater resilience, thereby enhancing the reliability of the classification process.

The research also compares two classifiers—Minimum Distance to Means (MDM) and k -Nearest Neighbors (k -NN)—both of which rely on distance metrics for classification. Chapter 5 provides an in-depth analysis of these classifiers.

To ascertain the practical utility of these classifiers and distance metrics, a controlled simulation study was conducted. Detailed in Chapter 7, this study evaluates the classifiers under ideal conditions where model assumptions are fully met. The simulation results facilitate a straightforward comparison of classifier performances, demonstrating the influence of different distance metrics on classification accuracy.

Finally, the methods are applied to real-world electroencephalography (EEG) data, as

detailed in Chapter 8. The analysis of EEG signals, with their complex spectral content, represents a challenging application domain. This application illustrates how theoretical advancements and findings from the simulation study are translated into practical tools for real-time data analysis. This thesis thus contributes to both the theoretical understanding of spectral distances and classification, and provides practical solutions for analyzing complex time series data in fields such as neuroscience.

Through these contributions, this research advances the field of time series classification by integrating geometric insights with practical applications, leading to more robust and accurate classification methods based on spectral density analysis.

Chapter 2

Time Series Theory

This section presents an overview of the fundamental concepts and properties of time series analysis. It is mainly inspired by the book [40].

A time series is a specific type of data that captures the dynamic evolution of a variable's value over time. In contrast to static data, time series data offers a sequential representation of the evolution of a specific variable at distinct points in time. This distinctive attribute renders time series data invaluable in a multitude of fields where the comprehension of temporal patterns and trends are of paramount importance.

In essence, a time series can be defined as a sequence of data points, typically measured at successive times, spaced at uniform intervals. The data points represent the values of a random variable observed and recorded over time. The chronological order of observations is what distinguishes time series data from other data types. This order preserves the time-dependent structure of the data, which is essential for analysis.

The prevalence of time series data across numerous disciplines reflects the importance of temporal analysis in understanding complex phenomena. To illustrate, in the field of economics, the evolution of the S&P 500 stock value is a time series that financial analysts closely examine in order to predict market trends, assess economic conditions, and inform investment decisions. Figure 2.0.1 illustrates the daily value fluctuations of the S&P 500 stock from May 2013 to August 2023, demonstrating the inherent variability and trends in financial markets.

In the field of biology, time series data are employed to monitor the population dynamics of various species, observe changes in the ecological environment, and investigate behavioral patterns. One illustrative example is the population evolution of polar bears, which is influenced by a range of factors including climate change and habitat loss. This time series is analyzed by ecologists to gain insights into species survival and inform conservation efforts.

Furthermore, time series analysis is not limited to single-variable studies. It is possible to work with multiple variables simultaneously, thereby facilitating the exploration of relationships and interactions between different phenomena. This multivariate approach enables the analysis of how the evolution of one variable influences or correlates with another. For example, by observing greenhouse gas emissions alongside polar bear populations in

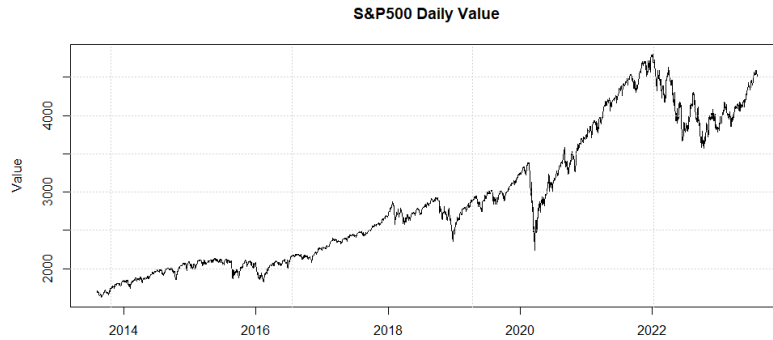


Figure 2.0.1: Evolution of the daily value of the S&P500 stock from May 2013 to August 2023.

a single time series, researchers can investigate the impact of environmental changes on wildlife. Such analyses are crucial for developing comprehensive models that consider the interplay of various factors over time.

To illustrate the concept of multivariate time series, let us consider a multivariate time series generated by a Vector Autoregressive (VAR(2)) process. A VAR(2) process is a statistical model that captures the linear interdependencies among multiple time series. Specifically, in a VAR(2) process ¹, the value of each variable at a given time is influenced by its own previous values and the previous values of all other variables in the system. This simulated 2D time series will be useful to clearly introduce and explain various concepts in a controlled and simple setting, providing a concrete example to better understand the theory. Figure 2.0.2 shows an example of this 2D time series of length $T = 1024$. The example has been simulated using the R-package `pdSpecEst` developed by the Joris Chau and explained in details in the PhD thesis [8]. It represents a bivariate VAR(2) based on Gaussian white noise.

Time series data also plays a critical role in fields like meteorology, where it is used to forecast weather patterns, and in engineering, where it helps monitor and predict system behaviors. The temporal dimension of time series data provides insights that are not accessible through static data analysis, making it an indispensable tool in scientific research and practical applications.

As we proceed with our investigation of time series theory, we will examine the mathematical foundations, statistical properties, and analytical techniques that underpin the study of time series. This foundation will equip us with the tools to uncover patterns and derive meaningful conclusions from time-dependent data. The subsequent sections will

¹More informations about this type of process can be found in the Chapter 7. The model used for this example is $\mathbf{X}(t) = \Phi_1 \mathbf{X}(t-1) + \Phi_2 \mathbf{X}(t-2) + \mathbf{Z}_t$, where $\Phi_1 = \begin{pmatrix} 0.3 & -0.5 \\ 0.5 & -0.7 \end{pmatrix}$, $\Phi_2 = \begin{pmatrix} -0.2 & 0.5 \\ -0.5 & 0.2 \end{pmatrix}$ and $\mathbf{Z}_t \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & 0.71 \\ 0.71 & 1 \end{pmatrix}$.

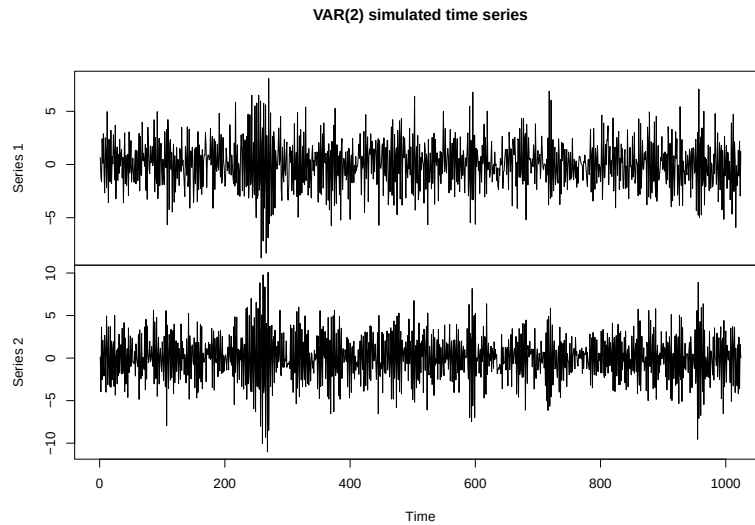


Figure 2.0.2: Example of a bivariate time series of length $T = 1024$ generated by a VAR(2) process.

cover definitions, key properties that form the backbone of time series analysis.

2.1 Definition and Statistical Properties

Time series analysis is a powerful statistical tool used to understand, model, and forecast data that is collected sequentially over time. This section aims to define what a time series is and describe its fundamental statistical properties. By delving into these definitions and properties, we can establish a solid foundation for more advanced analyses.

2.1.1 Definition of Time Series

A time series is formally defined as a sequence of data points collected or recorded at successive points in time. These data points are typically spaced at uniform intervals, although time series with irregular intervals also exist. Mathematically, a time series can be represented as follows:

Definition 2.1.1 (Time Series). *Let X be a random variable of dimension P and $\mathcal{T}$ be an ordered set. A time series $\{\mathbf{X}(t)\}_{t \in \mathcal{T}}$ is a sequence of observations of X at each time $t \in \mathcal{T}$. For the purposes of this thesis, the time series considered will be discrete and real-valued, which means that $\mathcal{T} = \mathbb{N}$ and X will be a real random variable.*

This definition encapsulates the essence of time series data: it is inherently ordered by time, and this ordering is crucial for any analysis. Unlike other types of data, where observations might be independent, time series data often exhibits dependencies and patterns over time.

The components of the vector, denoted $\mathbf{X}_i(t)$, represent the observation at time t for the i -th variable. For instance, if the first variable in the model represents the price of wheat over time, $y_1(1998)$ would denote the price of wheat in 1998.

2.1.2 Statistical Properties of Time Series

To analyze time series data effectively, we need to understand its statistical properties. These properties help us to describe the underlying structure and behavior of the data. The key statistical properties of multivariate time series include mean, variance, and autocovariance.

2.1.2.1 Mean

The mean of a time series provides a measure of its central tendency over time. For a time series $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ of dimension P , the mean is defined as:

$$\mu_X(t) = \left(\mathbb{E}(\mathbf{X}_1(t)), \dots, \mathbb{E}(\mathbf{X}_P(t)) \right)^T, \quad \forall t \in \mathbb{N}$$

2.1.2.2 Variance

The variance of a time series measures the dispersion of its values around the mean. It provides insight into the volatility of the data. For a time series $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ of dimension P , the variance-covariance matrix is defined as:

$$\sigma_X^2(t) = \begin{pmatrix} \text{var}(\mathbf{X}_1(t)) & \dots & \text{cov}(\mathbf{X}_1(t), \mathbf{X}_P(t)) \\ \vdots & \ddots & \vdots \\ \text{cov}(\mathbf{X}_P(t), \mathbf{X}_1(t)) & \dots & \text{var}(\mathbf{X}_P(t)) \end{pmatrix}, \quad \forall t \in \mathbb{N}$$

2.1.2.3 Autocovariance

While mean and variance are important, they do not capture the temporal dependencies in a time series. Autocovariance is a measure that helps to quantify the linear dependency between observations at different times between different dimensions. It is defined as:

Definition 2.1.2 (Autocovariance). *Let $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ be a time series of dimension P . The autocovariance function of X , denoted Γ_X , is defined by*

$$\Gamma_X(s, t) := \begin{pmatrix} \text{cov}(\mathbf{X}_1(s), \mathbf{X}_1(t)) & \dots & \text{cov}(\mathbf{X}_P(s), \mathbf{X}_1(t)) \\ \vdots & \ddots & \vdots \\ \text{cov}(\mathbf{X}_P(s), \mathbf{X}_1(t)) & \dots & \text{cov}(\mathbf{X}_P(s), \mathbf{X}_P(t)) \end{pmatrix}, \quad \forall s, t \in \mathbb{N}$$

The autocovariance function $\Gamma_X(s, t)$ provides a measure of how values at different times are related. On the diagonal of the matrix, this is the autocovariance of a dimension with itself, which means that this measures the linear dependency at different times of a particular dimension. The other elements measure the linear dependency at different times

between two dimensions. If the autocovariance is high at the time (s, t) , it indicates that the values at those times are strongly correlated.

2.1.2.4 Stationarity

A fundamental concept in time series analysis is stationarity. A time series is said to be stationary if its statistical properties do not change over time. Stationarity simplifies the analysis and modeling of time series data. There are different types of stationarity, but we will focus on weak stationarity.

Definition 2.1.3 (Weak Stationarity). *A time series $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ is weakly stationary if the following conditions are met:*

1. *The expected value is constant: $\mu_X(t) = \mu_X \quad \forall t \in \mathbb{N}$.*
2. *The variance is finite and constant: $\sigma_X^2(t) = \sigma_X^2 < \infty \quad \forall t \in \mathbb{N}$.*
3. *The autocovariance between two observations depends only on the time lag between them: $\Gamma_X(r, r+h) = \Gamma_X(s, s+h) \quad \forall h, r, s \in \mathbb{N}$.*

For weakly stationary time series, the autocovariance function $\Gamma_X(t, t+h)$ depends only on the time lag h , not on the actual times t and $t+h$. Therefore, we can simplify the notation by defining the autocovariance function based on the time lag:

$$R_X(h) := \Gamma_X(t, t+h) \quad \forall h, t \in \mathbb{N} \quad (2.1.1)$$

This autocovariance function has the following properties :

1. *Orthogonality : $R_X(h) = R_X(-h)^t$.*
2. *Cauchy-Schwarz Inequality :*

$$\left(R_X(h) \right)_{(i,j)} \leq \left(R_X(0) \right)_{(i,i)} \left(R_X(0) \right)_{(j,j)}, \quad \forall i, j = 1, \dots, P.$$

3. *Positive semi-definite : $\sum_{i,j=1}^T a_i^t R_X(i-j) a_j, \quad \forall T \geq 1$.*

Understanding and identifying stationarity in a time series is crucial because many analytical techniques and models assume or require the data to be stationary. It simplifies the complex dynamics within the data, making it more amenable to analysis, modeling, and forecasting. Non-stationary data can often be transformed into stationary data through various methods, such as differencing or detrending.

The example introduced at the beginning of the section, a VAR(2) process, is a case of a weakly stationary time series. This example helps illustrate the key statistical properties of such processes.

First, the conditions for weak stationarity are that the mean and variance of the time series do not change over time, and that the autocovariance between two observations

depends only on the time lag between them, not on the actual times at which they are measured. These properties simplify the analysis and modeling of the time series.

The first two conditions of weak stationarity can be quickly observed in Figure 2.0.2. The mean value remains constant over time, indicating no trend in the data. Similarly, the variance does not fluctuate, suggesting a stable level of dispersion around the mean. These visual confirmations are essential as they imply that the process does not exhibit long-term changes in its central tendency or variability, which are key indicators of stationarity.

The third condition, involving the autocovariance, is more complex but equally critical. In a stationary process, the autocovariance function should only depend on the time lag, not on the actual time points. This means that the relationship between two values in the time series is determined solely by how far apart they are in time, not by their position in the time series. This property allows us to define the autocovariance function based only on the time lag, as expressed in equation 2.1.1.

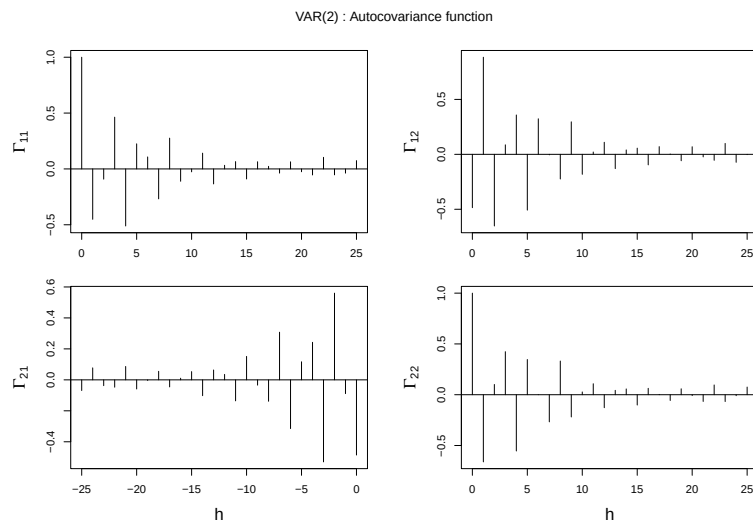


Figure 2.1.1: Autocovariance function of the example of VAR(2) process of the Figure 2.0.2.

Figure 2.1.1 illustrates the autocovariance function for our VAR(2) process. From the figure, we observe that the autocovariance is highest at smaller time lags and gradually decreases as the time lag increases. This behavior is typical in many real-world time series where observations close in time are more likely to be similar due to the influence of common factors or persistent effects. In the context of a VAR(2) process, this pattern can be explained by the fact that each observation is influenced by its own previous values as well as the previous values of other variables in the system, with a small random noise component. Therefore, the dependencies between observations that are close in time are strong, but as the time lag increases, these dependencies weaken.

Understanding the autocovariance function is crucial because it provides insight into the internal structure of the time series. The higher the autocovariance at a particular lag, the more the values at that lag are related. This information can be used to identify the

presence of cycles, trends, and other temporal structures in the data, which are essential for building accurate predictive models.

By defining and exploring these fundamental statistical properties, we lay the groundwork for more advanced time series analysis techniques. These properties not only help in describing the data but also in choosing appropriate models and making accurate predictions.

Chapter 3

Spectral Analysis of Time Series

In the study of time series, understanding the underlying structure and behavior of the data are crucial. Chapter 2 provided a comprehensive overview of the fundamental concepts and statistical properties of time series, including definitions, mean, variance, autocovariance, and stationarity. These properties form the foundation for analyzing temporal dependencies within time series data. However, to gain deeper insights, particularly in the frequency domain, additional tools are necessary.

The first step in any time series investigation always involves a careful examination of the recorded data plotted over time. This initial scrutiny often suggests the method of analysis and the statistics that will be useful for summarizing the information in the data. It is essential to recognize that two separate, but not necessarily mutually exclusive, approaches to time series analysis exist: the time domain approach and the frequency domain approach. The time domain approach focuses on investigating lagged relationships, asking questions such as "How does what happened today affect what will happen tomorrow?" In contrast, the frequency domain approach investigates cycles and periodicities within the data, such as economic cycles of expansion and recession. Both approaches offer valuable insights and are explored in details in the subsequent sections. However, in this master's thesis, we will focus on the frequency domain approach.

The spectral density function serves as a bridge between the time and frequency domains, offering a complementary perspective by analyzing how different frequency components contribute to the overall behavior of the time series. This frequency domain approach is particularly powerful for identifying and understanding periodicities and oscillatory behavior within the data. Applications of spectral density to signal classification have been well-established in various fields, including climatic data [14], protein sequences [20], and EEG signal classification [1, 28].

The global goal of this thesis is to develop effective methods for the classification of time series based on their spectral properties. By transforming time series data into the frequency domain and utilizing the spectral density function, we can extract features that are highly informative for classification tasks. This chapter delves into the theory of spectral density, exploring its definition, properties, and estimation methods. These concepts are pivotal for the subsequent development of classification algorithms that leverage spectral

information to distinguish between different types of time series.

3.1 Spectral Density

The spectral density function is an essential concept in the analysis of time series, especially in fields like signal processing, econometrics, and geophysics. It represents how the variance (or power) of a time series is distributed across different frequency components. By transforming the autocovariance function into the frequency domain, we can uncover periodicities and other structural features that are not immediately apparent in the time domain.

First, it is important to define what is meant by a Hermitian matrix, which plays a crucial role in the properties of spectral density matrices.

Definition 3.1.1 (Hermitian Matrix). *A $P \times P$ complex square matrix A is Hermitian if $A = A^*$, where A^* is the complex conjugate transpose of A . In other words,*

$$A = A^* \iff A_{ij} = \overline{A_{ji}} \quad \forall i, j = 1, \dots, P, \quad (3.1.1)$$

where $\overline{x}$ denotes the complex conjugate of x .

With this foundation, we can define the spectral density matrix, which is essential for understanding the frequency domain characteristics of time series.

Definition 3.1.2 (Spectral Density Matrix). *Let $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ be a discrete real-valued zero-mean weakly stationary time series of dimension P with an autocovariance function R_X satisfying,*

$$\sum_{j,k=1}^P \sum_{h=-\infty}^{+\infty} \left| \left(R_X(h) \right)_{(j,k)} \right| < \infty. \quad (3.1.2)$$

The $P \times P$ spectral density matrix of $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$, denoted f_X , is a function $f_X : [-\pi, \pi] \rightarrow \mathbb{C}^{P \times P} : \omega \mapsto f_X(\omega)$ and defined as the Fourier transform of the autocovariance function R_X , i.e.,

$$f_X(\omega) = (2\pi)^{-1} \sum_{h=-\infty}^{+\infty} R_X(h) \exp(-i\omega h), \quad \forall \omega \in [-\pi, \pi]. \quad (3.1.3)$$

We call "spectral matrices" the set of the values of the spectral density at different frequencies, i.e. the set $\{f_X(\omega) \mid \omega \in [-\pi, \pi]\}$. Moreover, we note $f_X^{(ii)}$ the element (i, i) of f_X , i.e.

$$f_X^{(ii)} : [-\pi, \pi] \rightarrow \mathbb{C} : \omega \mapsto f_X^{(ii)}(\omega) := \left(f_X(\omega) \right)_{(ii)} \quad (3.1.4)$$

The spectral density matrix $f_X(\omega)$ exhibits several important properties. It is (2π) -periodic, even, positive definite, and Hermitian for each frequency [4]. Mathematically, this can be expressed as:

$$f_X(\omega) = f_X(-\omega) = f_X^*(\omega), \quad \forall \omega \in [-\pi, \pi], \quad (3.1.5)$$

where $(*)$ denotes the complex conjugate transpose.

Consequently, thanks to the hermitian properties of the spectral density matrix, the auto-spectra are real and it suffices to consider only the elements $f_X^{(ij)}$ with $i \leq j$. Additionally, the evenness of the spectral density matrix enables us to restrict the frequency range from $[-\pi, \pi]$ to $[0, \pi]$. This reduction simplifies the analysis by focusing only on the necessary components and frequencies, thereby reducing computational complexity while preserving the essential information.

The spectral density matrix $f_X(\omega)$ is a powerful tool in time series analysis, as it encapsulates both the auto-spectral and cross-spectral information for each frequency. Specifically, the diagonal elements of the spectral density matrix, $f_X^{(ii)}(\omega)$, are known as the auto-spectra. They represent the distribution of power over frequency for the individual components of the time series. Essentially, the auto-spectrum $f_X^{(ii)}(\omega)$ of component $\mathbf{X}_i(t)$ reflects how much the variance of $\mathbf{X}_i(t)$ can be attributed to different frequency components.

The off-diagonal elements, $f_X^{(ij)}(\omega)$ for $i \neq j$, are known as the cross-spectra. The cross-spectrum $f_X^{(ij)}(\omega)$ measures the linear dependency between different components $\mathbf{X}_i(t)$ and $\mathbf{X}_j(t)$ at each frequency. This is important for understanding how different components of a multivariate time series interact with each other over different frequencies. If two components are highly correlated at a certain frequency, the cross-spectrum will have a significant magnitude at that frequency.

Understanding the auto-spectrum and cross-spectrum can provide deeper insights into the internal structure of a time series. For instance, the auto-spectrum can be seen as arising from the decomposition of the underlying process along sinusoidal waves with random amplitudes and phases. The cross-spectrum, on the other hand, captures the linear relationships between different time series components across various frequencies.

To illustrate the spectral density matrix more concretely, we will use the example introduced in the previous section, depicted in Figure 2.0.2. This example involves a bivariate time series modeled by a VAR(2) process. Consequently, the spectral density matrix for this process is a complex 2×2 matrix.

Due to the Hermitian property of the spectral density matrix, $f_X^{(ij)}(\omega) = \overline{f_X^{(ji)}(\omega)}$, we need to consider only the unique elements: the auto-spectra $f_X^{(11)}(\omega)$ and $f_X^{(22)}(\omega)$, and the cross-spectrum $f_X^{(12)}(\omega)$.

Figure 3.1.1 shows the spectral density matrix of the VAR(2) process across all frequencies from 0 to π . The diagonal plots illustrate the auto-spectra $f_{11}(\omega)$ and $f_{22}(\omega)$, representing the power distribution of each individual time series component over frequency. Significant peaks are observed at $\omega \approx 2.35$, indicating that this frequency component prominently contributes to the variance of both time series components.

The off-diagonal elements represent the cross-spectrum $f_{12}(\omega)$. The top right plot shows the real part of $f_{12}(\omega)$, while the bottom left plot shows the imaginary part. The cross-spectrum captures the linear dependency between the two time series components at each frequency. Significant features at $\omega \approx 2.35$ suggest strong interaction between the components at this frequency.

Peaks in the spectral density matrix indicate dominant periodic components within the

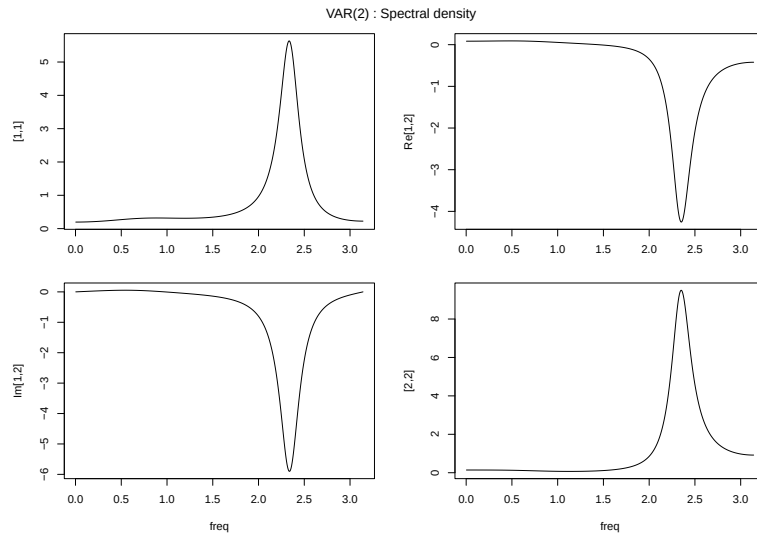


Figure 3.1.1: Spectral density matrix of the VAR(2) process example shown in Figure 2.0.2. The diagonal elements represent the auto-spectra f_{11} and f_{22} , while the off-diagonal elements represent the cross-spectrum f_{12} . The graph at the top right is the real part of the cross-spectrum f_{12} , and the graph at the bottom left is the imaginary part of the cross-spectrum f_{12} .

time series data. These peaks correspond to frequencies at which the power or variance of the signal is concentrated, revealing the primary oscillatory behavior of the process. Identifying these dominant frequencies is crucial for understanding the dynamic properties of the time series, as they reflect the inherent cycles or periodicities that characterize the data.

3.2 Estimator of the Spectral Density

In most cases, the true spectral density cannot be determined because the true value of the autocovariance function is unknown. Consequently, it is necessary to estimate it. One well-known nonparametric estimator is the periodogram. A nonparametric estimator is noteworthy because it allows us to make minimal assumptions about the underlying distribution of the data under study.

3.2.1 Periodogram

To estimate the spectral density matrix nonparametrically, one usually first converts the data from the time domain to the frequency domain using the discrete Fourier transform:

Definition 3.2.1 (Discrete Fourier Transform). *The discrete Fourier transform of a time*

series $\{X(t)\}_{t=1}^T$ of dimension P is defined as

$$\mathcal{F}_X(\omega_j) := T^{-1/2} \sum_{t=1}^T X(t) e^{-i\omega_j t}, \quad \text{for } j = -\frac{T}{2}, \dots, \frac{T}{2}, \quad (3.2.1)$$

where $\{\omega_j\}$ are the Fourier frequencies, i.e. $\omega_j = \frac{2\pi j}{T}$ and $j = -\frac{T}{2}, \dots, \frac{T}{2}$.

In practice, the discrete Fourier transform is quickly calculated (of order $O(T \log(T))$) via the famous FFT algorithm by Cooley and Tukey [11].

Definition 3.2.2 (Periodogram). *Let $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ be a time series of dimension P and length T . The periodogram of $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$, denoted I_X , is defined for all Fourier frequencies $\{\omega_j\}$ by:*

$$I_X(\omega_j) = (2\pi T)^{-1} \mathcal{F}_X(\omega_j) \mathcal{F}_X^*(\omega_j), \quad \text{for } j = -\frac{T}{2}, \dots, \frac{T}{2}, \quad (3.2.2)$$

where $(*)$ denotes the complex conjugate transpose and $\mathcal{F}_X$ is the discrete Fourier transform of $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$, defined by

$$\mathcal{F}_X(\omega) := T^{-1/2} \sum_{t=1}^T X(t) e^{-2\pi i \omega t}. \quad (3.2.3)$$

The asymptotic distribution of the periodogram of a time series $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ is given by [8, p. 12]:

$$I_X(\omega) \xrightarrow{d} \begin{cases} W_d^C(1, f_X(\omega)), & \text{if } \omega \neq 0 \pmod{\pi}, \\ W_d(1, f_X(\omega)), & \text{otherwise,} \end{cases} \quad \text{as } T \rightarrow \infty. \quad (3.2.4)$$

Here, $\xrightarrow{d}$ denotes convergence in distribution, and $W_d^C(1, f_X(\omega))$ and $W_d(1, f_X(\omega))$ are the complex and real-valued $(d \times d)$ -dimensional Wishart distributions with 1 degree of freedom and expected value $f_X(\omega)$, respectively. The Wishart distribution with 1 degree of freedom generalizes the chi-squared distribution to multivariate random variables.

Furthermore, the seminal book *Time Series: Data Analysis and Theory* by David Brillinger [4] presents detailed properties of the first and second moments of the periodogram. These properties are critical for understanding the estimation and variability of the periodogram.

Property 3.2.3 (Expectation of the Periodogram). *Let $\mathbf{X}(t)$ be a time series of dimension P , with $f_X(\omega)$ as its spectral density, and $I_X(\omega)$ as its periodogram. The expectation of the periodogram is*

$$\mathbb{E}[I_X(\omega)] = f_X(\omega) + O(T^{-1}), \quad (3.2.5)$$

where $O(\cdot)$ denotes the Big O notation [25].

Property 3.2.4 (Variance of the Periodogram). *Let $\mathbf{X}(t)$ be a time series of dimension P , $f_X(\omega)$ its spectral density, and $I_X(\omega)$ its periodogram. The variance of the periodogram is*

$$\mathbb{V} \left[\left(I_X(\omega) \right)_{i,j} \right] = \left(f_X(\omega) \right)_{i,j}^2 + \left(f_X(\omega) \right)_{i,i} \left(f_X(\omega) \right)_{j,j} + O(T^{-1}), \quad (3.2.6)$$

where $O(\cdot)$ denotes the Big O notation.

Property 3.2.5 (Cross-Frequency Covariance of the Periodogram). *Let $\mathbf{X}(t)$ be a time series of dimension P , $f_X(\omega)$ its spectral density, and $I_X(\omega)$ its periodogram. Let $\omega_i, \omega_j \in [-\pi, \pi]$, with $\omega_i \neq \omega_j$ be two distinct frequencies and $k_i, l_i = 1, \dots, P$. The cross-frequency covariance of the periodogram is approximately*

$$\text{Cov} \left[\left(I_X(\omega) \right)_{l_1, k_1}(\omega_i), \left(I_X(\omega) \right)_{l_2, k_2}(\omega_j) \right] \approx O(T^{-1}), \quad (3.2.7)$$

where $O(\cdot)$ denotes the Big O notation.

Although the periodogram is an asymptotically unbiased estimator of the spectral density, these properties imply that it is not a consistent estimator. As $T \rightarrow \infty$, we have:

$$\mathbb{E}[I_X(\omega)] \rightarrow f_X(\omega) \quad \text{and} \quad \mathbb{V} \left[\left(I_X(\omega) \right)_{i,j} \right] \rightarrow \left(f_X(\omega) \right)_{i,j}^2 + \left(f_X(\omega) \right)_{i,i} \left(f_X(\omega) \right)_{j,j}.$$

This indicates that while the expectation converges to the spectral density, the variance does not tend to zero. Moreover, the periodogram is not positive definite for time series with dimension $d > 1$ because it has rank one by construction. This rank deficiency further complicates the use of periodograms for higher-dimensional time series.

To illustrate this, we consider the VAR(2) process example introduced earlier. Figure 3.2.1 shows the periodogram estimation of the spectral density matrix for this example. The true spectral density matrix is depicted in red, whereas the periodogram is shown in gray. The periodogram is computed using the precious function `pdPgram()` from the package `pdSpecEst`[7]. We can observe that the periodogram captures the overall shape of the spectral density, with notable peaks at the correct frequencies. However, the periodogram exhibits high variability and instability, highlighting the need for more robust estimation techniques.

The above result highlights a significant problem with the periodogram: it is not a consistent estimator of the spectral density, $f_X(\omega)$. Since the periodogram converges to a nondegenerate Wishart distribution, it fails to provide a reliable estimate of the spectral density. Additionally, for multivariate time series ($d > 1$), the periodogram is not positive definite due to its inherent rank-one structure.

To address these issues, we can apply smoothing techniques to the periodogram. If the spectral density is assumed to be sufficiently smooth across frequency (e.g., continuously second-order differentiable), several common nonparametric procedures can yield consistent, positive definite estimates of the spectral density. One such method is smoothing the periodogram by applying locally weighted averaging across frequencies. Another approach involves averaging multiple tapered periodogram estimates, known as multitaper spectral estimation. These techniques help in mitigating the variability and rank-deficiency of the periodogram, providing more reliable spectral estimates.

3.2.2 Smoothed Periodogram

As previously stated, the periodogram is a fundamental tool for spectral analysis; however, it does not provide a consistent estimator due to its high variance. This inconsistency

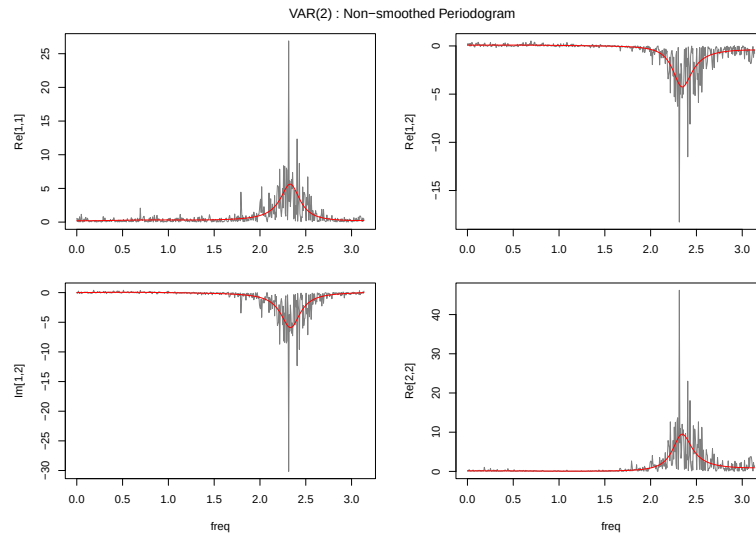


Figure 3.2.1: Estimation of the spectral density matrix shown in Figure 3.1.1 by the periodogram estimator. The true spectral density matrix is illustrated in red and the periodogram in gray.

presents a significant challenge when attempting to classify time series based on their spectral properties. The raw periodogram may exhibit substantial fluctuations, which may obscure the true underlying spectral features of the time series, thereby leading to unreliable comparisons between different series.

In the context of time series classification, the generation of accurate and stable estimates of spectral densities is of paramount importance. In the context of time series classification, the use of distance-based methods such as the k-Nearest Neighbors or Minimum Distance to Means is contingent upon the accuracy of the spectral properties being compared. The inherent noise in unsmoothed periodograms can lead to misclassification, particularly when the spectral properties of the time series are not clearly delineated. The high variance inherent in the periodogram can result in the misleading representation of spectral characteristics, which in turn can negatively impact classification accuracy.

It could be argued that the process of comparison itself—by aggregating information across frequencies—could mitigate some of this noise. However, studies have shown that proper smoothing of the periodogram significantly enhances the performance of time series classification algorithms. The application of a smoothing technique serves to stabilize the periodogram, thereby reducing the variance and facilitating the identification of the dominant spectral features. This enables a more accurate distinction between different time series based on their spectral properties.

To this end, two principal smoothing techniques are typically employed: *Kernel Smoothing* and *Multitapering Smoothing*. These methodologies are devised to attenuate the noise present in the periodogram while maintaining the salient spectral characteristics, thereby furnishing a more robust foundation for time series classification.

In this thesis, we will utilize the multitapering smoothing technique, which has demonstrated excellent performance and appears to be well-suited for the estimation task.

Kernel Smoothing : A popular nonparametric approach is using a smoothing kernel to dilute the information in a given window of frequencies [39].

Definition 3.2.6 (Kernel-Smoothed Periodogram). *Let $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ be a time series of dimension P and length T and I_X its periodogram. Let $K^{(ij)}$ be a smoothing kernel¹ for all $i, j = 1, \dots, P$ and $m \leq T/2$ be the smoothing span. The kernel-smoothed periodogram is defined for each element (i, j) at the Fourier frequencies $\{\omega_l\}$ by*

$$\tilde{f}_X^{(ij)}(\omega_l) = \sum_{k=-m}^m K^{(ij)}(\omega_l - \omega_{l+k}) I_X^{(ij)}(\omega_{l+k}), \quad l = -\frac{T}{2}, \dots, \frac{T}{2} \quad (3.2.8)$$

Usually, the most frequent choice of kernel is the rectangular smoothing kernel, which is defined as $K(\omega) = \frac{1}{2m+1}$ where m is the smoothing span.

This approach provides a consistent estimator of the spectral density matrix with a sufficient large frequency window. However, the larger the frequency window is, the slower the bias convergence will be.

Multitapering : Another approach to smooth the periodogram is the multitaper spectral estimation. This approach was developed by David J. Thomson in 1982 [45]. This method involves averaging multiple periodograms, each computed from a different tapered version of the time series. The tapering functions, also known as data tapers or windows, reduce the variance of the spectral estimate by mitigating the effects of spectral leakage. Spectral leakage occurs when the finite length of the data causes energy from a signal at one frequency to spill into neighboring frequencies, distorting the spectral estimate.

Let $\{h_k(t), t = 1, \dots, T\}$ for $k = 1, \dots, L$ be L orthogonal tapering functions, such that:

$$\sum_{t=1}^T h_k(t)^2 = 1, \quad \text{and} \quad \sum_{t=1}^T h_k(t) h_j(t) = 0, \quad \text{for } k \neq j.$$

We can define $\mathcal{F}_{h_k, X}(\omega)$ the discrete Fourier transform of the observed time series $\{X(t)\}$ tapered with the tapering function $h_k(t)$. That is,

$$\mathcal{F}_{h_k, X}(\omega_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^T h_k(t) X(t) e^{-it\omega_j}, \quad (3.2.9)$$

where $\{\omega_j\}$ are the Fourier frequencies, i.e. $\omega_j = \frac{2\pi j}{T}$ and $j = -\frac{T}{2}, \dots, \frac{T}{2}$.

Definition 3.2.7 (Multitaper Spectral Estimator). *Let $\{\mathbf{X}(t)\}_{t \in \mathbb{N}}$ be a time series of dimension P and length T , and I_X its periodogram. Let $\{h_k(t), t = 1, \dots, T\}$ for $k = 1, \dots, L$*

¹A kernel is a function $K : \mathbb{R} \rightarrow \mathbb{R}$ such that $\forall x \in \mathbb{R} \ K(x) \geq 0$, $K(-x) = K(x)$ and $\int_{\mathbb{R}} K(x) dx = 1$.

be L orthogonal tapering functions. The multitaper spectral estimator is defined for each element (i, j) at the Fourier frequencies $\{\omega_j\}$ by

$$\hat{f}_X^{(ij)}(\omega_j) = \frac{1}{L} \sum_{k=1}^L \mathcal{F}_{h_k, X}(\omega_j) \mathcal{F}_{h_k, X}(\omega_j)^*, \quad (3.2.10)$$

where $(*)$ denotes the complex conjugate transpose.

The multitaper method addresses the issues associated with periodograms by obtaining multiple independent estimates from the same sample using orthogonal tapers. Each taper is multiplied by the signal to create windowed trials, and the power at each frequency component is estimated. The final spectral density is obtained by averaging over all tapered spectra, thus recovering information lost due to signal attenuation from individual tapers.

This method is particularly useful with a small number of trials, reducing estimator variance more effectively than single taper methods. Even with many trials, multitapering provides better control over the trade-off between bias and variance.

A popular choice of orthogonal tapering functions is the DPSS (discrete prolate spheroidal sequence) or Slepian tapers, as first proposed by David Slepian [42]. The zeroth-order DPSS taper maximizes the spectral concentration within a given frequency band. Specifically, for a given B and $0 < W < \pi$, among all sequences $\{h(t), t = 1, \dots, B\}$, the zeroth-order DPSS sequence, with time-bandwidth parameter $n_W = TW$, maximizes the relative spectral energy $\lambda(B, W)$ in the frequency band $[-W, W]$ given by:

$$\lambda(B, W) = \frac{\int_{-W}^W \left| \sum_{t=1}^B h(t) e^{-it\omega} \right|^2 d\omega}{\int_{-\pi}^{\pi} \left| \sum_{t=1}^T h(t) e^{-it\omega} \right|^2 d\omega}.$$

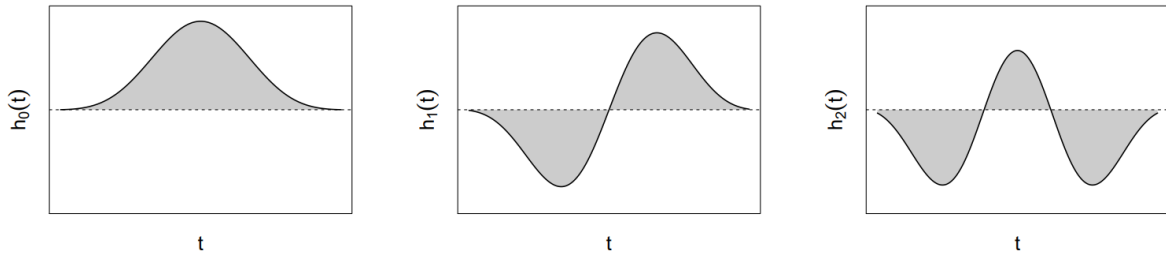


Figure 3.2.2: Three orthogonal DPSS taper functions $\{h_k(t), t = 1, \dots, B\}$ of order $k = 0, 1, 2$ in the time domain, with time-bandwidth parameter $n_w = 3$. [8]

These DPSS functions are chosen due to their orthogonality and ability to minimize spectral leakage, which is particularly useful when dealing with a small number of trials, as they effectively reduce estimator variance. Even with larger datasets, multitapering offers better control over the trade-off between bias and variance.

In practice, a weighted average compensates for energy loss at higher-order tapers, ensuring a stable and reliable spectral density estimate, especially with small datasets or

when high precision is required. Multitaper spectral estimation thus provides a consistent spectral density estimate with reduced variance compared to the raw periodogram, making it crucial for accurate spectral analysis, particularly in high-dimensional time series data.

To illustrate the multitaper method, we revisit the VAR(2) process example introduced earlier. Figure 3.2.3 illustrates the multitaper spectral estimator of the spectral density matrix for this example. The true spectral density matrix is depicted in red, while the periodogram is shown in gray. The multitaper spectral estimator is computed using the `pdPgram()` function from the `pdSpecEst` package [7].

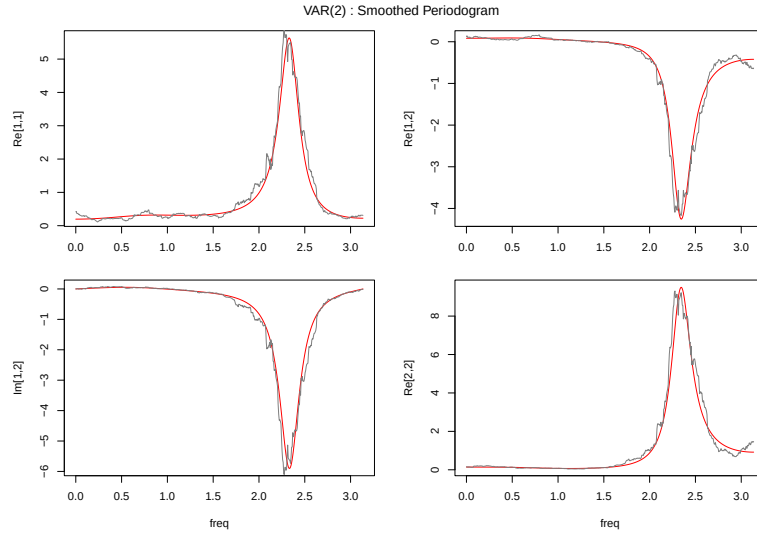


Figure 3.2.3: Estimation of the spectral density matrix shown in Figure 3.1.1 using the multitaper spectral estimator. The smoothing parameters chosen are $B = 20$ and $n_w = 30$. The true spectral density matrix is illustrated in red, and the multitaper spectral estimator in gray.

As already mentioned, the multitaper method involves averaging multiple tapered periodograms to produce a more stable estimate of the spectral density matrix. By using orthogonal tapers, the estimator reduces the variance while retaining a high resolution. In this example, the smoothing parameters $B = 20$ and $n_w = 30$ were chosen to balance bias and variance effectively.

From Figure 3.2.3, we observe that the multitaper spectral estimator closely follows the true spectral density matrix, significantly improving upon the periodogram estimator. The periodogram, depicted in Figure 3.2.1, exhibits high variability and is prone to large fluctuations, making it less reliable for spectral density estimation. In contrast, the multitaper spectral estimator provides a smoother and more accurate approximation.

This enhanced accuracy of the multitaper spectral estimator is particularly evident when comparing it to the true spectral density matrix (red curve). The reduced variance of the multitaper spectral estimator implies that it is less affected by the inherent noise in the data, leading to a more precise estimation. This observation suggests that using the

multitaper spectral estimator for comparing time series could be more reliable than using periodograms, which tend to be overly variable.

Chapter 4

Geometry of Spectral Density

In the previous chapter, we delved into the spectral theory of time series, focusing on the spectral density function and its estimation. These spectral properties provide a powerful means of analyzing time series data in the frequency domain. As we move forward in our journey to classify time series based on their spectral characteristics, it is essential to understand the underlying geometric structure of spectral density matrices.

This chapter explores the geometry of spectral density matrices, which is crucial for effectively comparing and classifying them. By leveraging the tools and concepts from differential geometry, we can utilize more sophisticated and accurate methods for measuring distances between spectral density matrices. This geometric approach allows us to make more informed decisions in the classification process, enhancing the accuracy and reliability of our methods.

Drawing from foundational concepts in Petersen's "Riemannian Geometry" [38] and Lee's "Introduction to Smooth Manifolds" [26], this chapter provides a rigorous theoretical basis for understanding the geometric aspects of spectral density matrices. These concepts are vital for developing advanced classification algorithms that exploit the geometric properties of the spectral domain.

To illustrate the importance of geometry in understanding distances, let us consider a simple analogy involving the Earth. Imagine we want to calculate the distance between two friends, one living in Louvain-La-Neuve and the other in New York, as illustrated in the Figure 4.0.1. While all objects on Earth exist in a three-dimensional (3D) space, they are confined to the Earth's surface. In this 3D world, we could calculate the distance between these two friends using the classical Euclidean distance. However, it is more relevant and intuitive to use the geographical or geodesic distance, which measures the distance along the Earth's surface [33].

This example highlights the importance of choosing an appropriate geometric framework for measuring distances. Similarly, understanding the geometry of spectral density matrices enables us to compare them effectively, ensuring that our classification methods are both accurate and meaningful. The next chapter will build on this foundation, focusing on practical applications and developing robust classification algorithms that utilize the geometric properties of spectral density matrices.

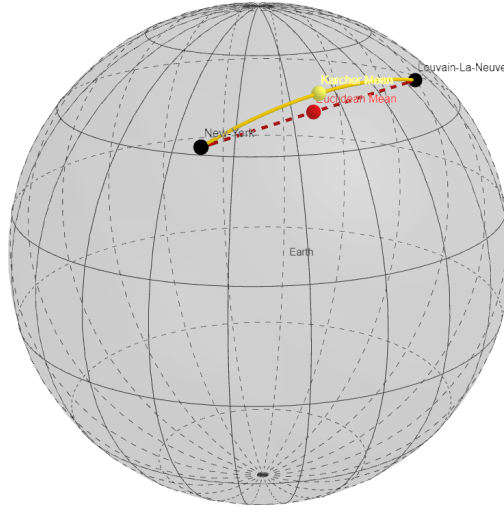


Figure 4.0.1: Distances between Louvain-La-Neuve and New York. The dashed red line inside the sphere represents the three-dimensional Euclidean distance. The yellow line represents the geodesic distance. The Euclidean and Karcher Means are represented by the red and yellow dots, respectively.

In summary, this chapter bridges the spectral analysis methods discussed previously and the practical classification techniques that follow. By understanding the geometry of spectral density matrices, we set the stage for more advanced and precise classification methodologies, ultimately achieving the global goal of this thesis: effective classification of time series based on their spectral densities.

4.1 Theoretical Foundations of Differential Geometry

4.1.1 Manifolds

To intuitively understand the concept of a manifold, let us consider it as a mathematical space that locally looks like Euclidean space. For instance, the surface of the Earth, though globally curved, appears flat when observed on a small scale, like standing in a small field. This local flatness allows us to apply familiar mathematical tools from calculus and geometry, which are typically used in flat, Euclidean spaces, to study manifolds.

We can define a manifold more formally in the following way [8]:

Definition 4.1.1 (Manifold). *A manifold $\mathcal{M}$ is a topological space that locally resembles Euclidean space. Formally, $\mathcal{M}$ is a topological manifold of dimension n if it satisfies the following properties:*

- Hausdorff property: *Any two distinct points $x, y \in \mathcal{M}$ are neighborhood-separable,*

i.e., there exist disjoint open subsets $U, V \in \mathcal{M}$ with $U \cap V = \emptyset$ such that $x \in U$ and $y \in V$.

- Second-countable: $\mathcal{M}$ has a countable basis for its topology.
- Locally Euclidean of dimension n : Each point in $\mathcal{M}$ has a neighborhood $U \subseteq \mathcal{M}$ that is homeomorphic to an open subset $V \subseteq \mathbb{R}^n$, where a homeomorphism $\varphi : U \rightarrow V$ is a function that is bijective, continuous and with a continuous inverse.

More information about the definitions of manifolds can be found in [26, 8].

In this thesis, we will work exclusively with smooth or C^∞ manifolds, which are topological manifolds endowed with an additional structure allowing us to define smooth functions (i.e., functions that are infinitely differentiable) on the manifold. Specifically, the definition of a smooth manifold relies on smooth maps between Euclidean spaces using coordinate charts. If U and V are open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then a map $f : U \rightarrow V$ is said to be smooth or C^∞ if each component function of f has continuous partial derivatives of all orders. If the smooth map f is also a homeomorphism, it is called a diffeomorphism.

4.1.2 Riemannian Geometry

Riemannian geometry is the study of smooth manifolds equipped with a Riemannian metric, which allows us to measure distances and angles on the manifold. A Riemannian manifold (M, g) consists of a smooth manifold M and a Riemannian metric g , a smoothly varying positive-definite inner product on the tangent space at each point.

The Riemannian metric will allow us to calculate the distance between two elements of a manifold. Intuitively, the distance between two points $p, q \in \mathcal{M}$ is the shortest path on the manifold $\mathcal{M}$ that connects the two points. Therefore, we need to define what is a path, or a curve, on a manifold.

Definition 4.1.2 (Curve). *A curve on a manifold M is a smooth map $\gamma : [a, b] \rightarrow M$. If $\gamma(a) = p$ and $\gamma(b) = q$, the curve connects the points p and q . The length of a curve γ with respect to the Riemannian metric g is given by:*

$$L(\gamma) = \int_a^b \sqrt{g(\gamma'(t), \gamma'(t))} dt. \quad (4.1.1)$$

Now, we can define the distance between two points on a manifold.

Definition 4.1.3 (Riemannian distance). *The distance $d(p, q)$ between two points $p, q \in M$ is the infimum of the lengths of all curves connecting p and q :*

$$d(p, q) = \inf_{\gamma} L(\gamma), \quad (4.1.2)$$

where γ ranges over all curves connecting p and q .

The shortest curve is called a geodesic. Geodesics generalize the concept of "straight lines" to curved spaces and are critical in defining distances on manifolds.

While the detailed mathematical framework presented in this section, including the rigorous definitions of manifolds, curves, and Riemannian distances, provides a robust foundation for understanding the geometry of spectral density matrices, not all these concepts will be directly employed in the subsequent sections of this thesis. The primary purpose of this discussion is to offer the reader a comprehensive background, ensuring that the core ideas are well-understood. Moving forward, our focus will be more applied: we will primarily leverage the concepts of manifolds, the Riemannian metric, and the Karcher mean, which is the mean of points on a manifold (c.f section 4.2.1), to develop and analyze classification methods for time series in the frequency domain. These concepts will serve as the mathematical tools necessary for defining distances between spectral densities and for constructing representative means of spectral densities, which are crucial for the classification tasks at the heart of this thesis.

4.2 Spectral Density Matrices and Manifolds

As discussed previously, $P \times P$ spectral matrices have particular properties. They are Hermitian and positive definite. Then, we can define the following subsets of the set of the $P \times P$ complex matrices, $\mathbb{C}^{P \times P}$.

Definition 4.2.1 (Hermitian Matrices Set). *The Hermitian matrices set is a subset of $P \times P$ complex matrices, $\mathbb{C}^{P \times P}$ and is defined by*

$$\mathcal{H}_P := \left\{ A \in \mathbb{C}^{P \times P} \mid A = A^* \right\}, \quad (4.2.1)$$

where A^* denotes the complex conjugate transpose of A .

Definition 4.2.2 ((HPD) Hermitian and Positive Definite Matrices Set). *The HPD (Hermitian and Positive Definite) matrices set is a subset of $P \times P$ complex matrices, $\mathbb{C}^{P \times P}$ and is defined by*

$$\mathcal{H}_P^+ := \left\{ A \in \mathbb{C}^{P \times P} \mid A = A^*, A \succeq 0 \right\}, \quad (4.2.2)$$

where A^* denotes the complex conjugate transpose of A .

Let $\mathbb{C}^{P \times P}$ denotes the set of $P \times P$ complex matrices. We can further define $\mathcal{H}_P$ as the set of Hermitian matrices and $\mathcal{H}_P^+$ as the set of Hermitian and positive definite matrices. Trivially, the HPD matrices set $\mathcal{H}_P^+$ is a subset of the Hermitian matrices set $\mathcal{H}$. Moreover, [29, 8] shows the property that the $\mathcal{H}_P^+$ is a manifold in the real linear vector space $\mathcal{H}$. Due to the Hermitian constraint, the dimension of this manifold is P^2 .¹

¹Intuitively, let us consider a Hermitian matrix A . By definition, the entries on the diagonal of A must be real numbers. Since there are P diagonal entries, this contributes P degrees of freedom. Next, let us consider the upper triangular portion of A (excluding the diagonal). This portion contains $\frac{P(P-1)}{2}$ entries, each of which is a complex number. Each complex number can be represented by two real numbers (its real and imaginary parts), contributing $2 \times \frac{P(P-1)}{2} = P(P-1)$ additional degrees of freedom. The lower

This property is crucial because, as mentioned above, spectral matrices are HPD, so they belong to the $\mathcal{H}_P^+$ manifold. Consequently, they can inherit all the properties inherent in manifold theory.

However, they do not form a Euclidean space (vector space with inner product) as HPD matrices do not form a vector space. This is because it does not constitute a vector space—they are not closed under addition or scalar multiplication. Consequently, utilizing Euclidean geometry tools for spectral density matrices is neither natural nor adequate.

Nevertheless, classification algorithms necessitate a notion of distance between different spectral density matrices. Although one might be inclined to use the Euclidean distance for this purpose, it is not an appropriate metric for spectral density matrices. Instead, it is essential to employ distances that consider the Riemannian structure of the space of spectral density matrices. This approach ensures that the inherent geometric properties of these matrices are respected, leading to more meaningful and accurate comparisons.

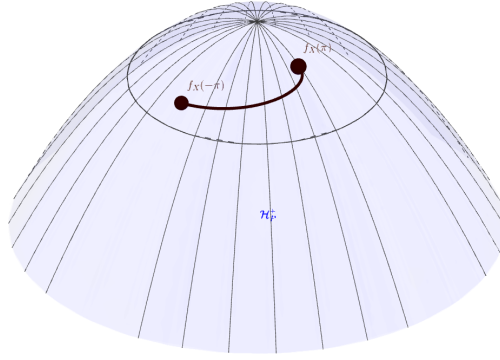


Figure 4.2.1: Representation of the spectral density matrix $f_X(\omega)$ from the frequency ω_1 to ω_2 as a curve in the manifold $\mathcal{H}_P^+$. The Figure shows an intuitive visualization, i.e., that is not the real manifold.

Given that the spectral matrices belong to the manifold $\mathcal{H}_P^+$, the spectral density matrix $f_X(\cdot)$ can be seen as a function

$$f_X : [-\pi, \pi] \rightarrow \mathcal{H}_P^+.$$

By observing Definition 4.1.2, we see directly that the spectral density matrix $f_X(\cdot)$ is a curve in $\mathcal{H}_P^+$. Figure 4.2.1 illustrates the representation of a spectral density as a curve in the manifold $\mathcal{H}_P^+$. Therefore, comparing spectral densities involves comparing these curves, implying that the analysis directly engages with the geometrical structures of the spectral density matrices.

triangular portion of A is not independent because, in a Hermitian matrix, it is the conjugate transpose of the upper triangular portion. Therefore, it does not add any further degrees of freedom.

In total, the degrees of freedom of A sum to $P + P(P - 1) = P^2$. This explains intuitively why the dimension of the manifold $\mathcal{H}_P^+$ is P^2 .

4.2.1 Karcher Mean

The Karcher or Fréchet mean offers a robust method to compute the "average" of points on a manifold, respecting its intrinsic geometric structure. This is particularly important for spectral density matrices, which reside on a curved manifold rather than in Euclidean space. In the context of classification, the Karcher mean can serve as a representative or "center" of a class, providing a meaningful way to summarize and compare sets of spectral density matrices.

To understand the concept intuitively, let us consider averaging points on the Earth's surface. Naively computing the arithmetic mean of their latitude and longitude coordinates might yield a point that does not lie on the Earth's surface. Instead, we should find a point that minimizes the total squared geodesic distances to all other points on the surface. This point is the Karcher mean.

Similarly, spectral density matrices do not reside in a flat space where simple arithmetic means are appropriate. Instead, they lie on a manifold where the Karcher mean provides a meaningful average by minimizing the sum of squared distances defined by the manifold's geometry.

Definition 4.2.3 (Karcher mean). *Let $\{A_i\}_{i=1,\dots,K}$ be a set of K points on a manifold $\mathcal{M}$ and d a given distance. The Karcher mean $G_d(A_1, \dots, A_K)$ is defined as the point on $\mathcal{M}$ that minimizes the sum of squared distances to all the points A_i :*

$$G_d(A_1, \dots, A_K) := \arg \min_{A \in \mathcal{M}} \sum_i d^2(A, A_i). \quad (4.2.3)$$

In classification tasks, the Karcher mean can be used to compute the mean spectral density matrix for each class. This mean matrix serves as a representative of the class, enabling the use of distance-based classification methods such as the Minimum Distance to Means (MDM) that will be defined in the Chapter 5. By comparing a new observation to the Karcher means of different classes, we can classify the observation based on the closest mean, thereby leveraging the geometric properties of the manifold for more accurate classification.

Computing the Karcher mean with a distance d involves iterative optimization techniques due to the lack of a closed-form solution. Starting from an initial guess A_0 , we iteratively update the guess by moving along the gradient of the objective function $\sum_i d^2(A, A_i)$. This process ensures convergence to the local minimum, which is the Karcher mean.

The utility of the Karcher mean associated with the Riemannian distance, as defined in 4.1.3, is significant in classification. By providing a robust and geometrically meaningful average of spectral density matrices, it enhances the accuracy of classification algorithms. The Karcher mean respects the intrinsic structure of the data, making it a powerful tool for summarizing and comparing spectral density matrices in a manner that simple arithmetic means cannot achieve. This geometric approach ensures that the classification process is both accurate and meaningful, leveraging the full potential of spectral and geometric properties of the data.

The desired properties for a mean in the context of spectral density matrices and other applications can be guided by the ALM properties list introduced by Ando et al. [3], which includes:

1. **Joint homogeneity:** $G(\alpha_1 A_1, \dots, \alpha_K A_K) = (\alpha_1 \cdot \dots \cdot \alpha_K)^{1/K} G(A_1, \dots, A_K)$.
2. **Permutation invariance:** For any permutation $\pi(A_1, \dots, A_K)$ of $(A_1, \dots, A_K)$,

$$G(A_1, \dots, A_K) = G(\pi(A_1, \dots, A_K)).$$
3. **Monotonicity:** If $A_i \geq B_i$ for all i , then $G(A_1, \dots, A_K) \geq G(B_1, \dots, B_K)$ in the positive semidefinite ordering.
4. **Continuity from above:** If $\{A_1^{(n)}\}, \dots, \{A_K^{(n)}\}$ are monotonic decreasing sequences (in the positive semidefinite ordering) converging to $A_1, \dots, A_K$, respectively, then $G(A_1^{(n)}, \dots, A_K^{(n)})$ converges to $G(A_1, \dots, A_K)$.
5. **Congruence invariance:** $G(S^T A_1 S, \dots, S^T A_K S) = S^T G(A_1, \dots, A_K) S$ for any invertible S .
6. **Joint concavity:** $G(\lambda A_1 + (1 - \lambda) B_1, \dots, \lambda A_K + (1 - \lambda) B_K) \geq \lambda G(A_1, \dots, A_K) + (1 - \lambda) G(B_1, \dots, B_K)$.
7. **Invariance under inversion:** $G(A_1, \dots, A_K)^{-1} = G(A_1^{-1}, \dots, A_K^{-1})$.
8. **Determinant identity:** $\det G(A_1, \dots, A_K) = (\det A_1 \cdot \dots \cdot \det A_K)^{1/K}$.

These properties are crucial in various applications, such as medical imaging, signal processing, and clustering methods, where finding a central representative for a collection of symmetric positive definite (SPD) matrices is essential. For example, the Karcher mean is valuable in aggregating several noisy measurements of the same object or determining a cluster center in classification tasks. There may be many different definitions of the means that satisfy all the properties we will discuss in the next chapter.

A significant issue in computing the arithmetic mean of SPD matrices is the *swelling effect*, which refers to the increase of the matrix determinant after averaging. This phenomenon can distort the true central tendency of the data, leading to incorrect conclusions in applications such as diffusion tensor imaging [13, 49]. The determinant of a matrix intuitively represents the volume of the ellipsoid described by the covariance matrix, with its axes pointing in the direction of the eigenvectors and lengths proportional to the square roots of the eigenvalues.

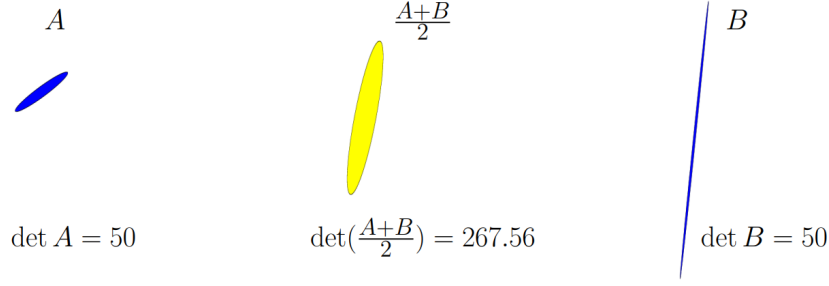


Figure 4.2.2: Illustration of the swelling effect when computing the arithmetic mean of two SPD matrices A and B . The determinants of A and B are both 50, while the determinant of their arithmetic mean $\frac{A+B}{2}$ is 267.56.

4.2.1.1 Karcher and Euclidean Mean comparison

Consider a set of three points $S_1, S_2, S_3 \in \mathcal{H}_P^+$. In Euclidean space, their classical mean would be computed as:

$$S_{\text{mean}} = \frac{1}{3}(S_1 + S_2 + S_3).$$

However, this simple averaging does not account for the manifold's curvature and might result in a matrix that is not positive definite or even Hermitian.

Instead, by computing the Karcher mean, we respect the manifold's structure. Assume S_1, S_2 , and S_3 lie on a curved manifold $\mathcal{H}_P^+$. The Karcher mean $\bar{S}$ is found by minimizing the sum of squared Riemannian distances:

$$\bar{S} = \arg \min_{S \in \mathcal{H}_P^+} \left(d_R^2(S, S_1) + d_R^2(S, S_2) + d_R^2(S, S_3) \right).$$

This approach ensures that $\bar{S}$ is a valid spectral density matrix on the manifold $\mathcal{H}_P^+$.

To illustrate, let us visualize the different means using the basic example used in the beginning of the section in the Figure 4.0.1. The Euclidean mean lies off the surface, whereas the Karcher mean lies on the surface, preserving the manifold's properties.

In conclusion, the Karcher mean provides a powerful and geometrically consistent method for averaging points on a manifold, essential for the accurate analysis and comparison of spectral density matrices.

Chapter 5

Method of classification

The classification of time series is a well explored field, involving a large number of different methods. Some of them are based directly on the time series values, others with transformed time series. For example, using a Deep Neural Networks [21], Dynamic Time Warping (DTW) distance between time series values [30], or using the covariance matrix of the filtered time series at some frequencies [46]. In this thesis, we will work in the frequency domain of time series, i.e. using spectral densities. For this, the time series must be discrete, real-valued, weakly stationary and have an absolutely summable autocovariance function. We will therefore consider that we deal with such time series.

Let us describe the general problematic of this thesis. Suppose that we aim to classify N time series $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$ of length T , and dimension P , into L classes. We denote $f_{\mathbf{X}_1}, f_{\mathbf{X}_2}, \dots, f_{\mathbf{X}_N}$ the spectral densities of the respective observed time series and $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_L$ the different classes. Moreover, we assume that each class represents one underlying process, which means that for each class $\mathcal{C}_l$ there is a true spectral density denoted ϕ_l .

As the spectral densities are usually not known, we estimate the spectral densities $f_{\mathbf{X}_i}$ using the multitaper spectral estimator $\hat{f}_{\mathbf{X}_i}$, defined previously in Definition 3.2.7.

In this chapter, two methods will be introduced and analysed: the *k-Nearest Neighbors* and the *Minimum Distance to Means*. Both methods are based on a distance to compare the time series and more precisely two compare their estimated spectral densities. The choice of such a distance will be discussed later in this thesis.

5.1 k-Nearest Neighbors

Let us introduce the first algorithm we will utilize in this thesis: the *k-Nearest Neighbors* (*k-NN*). The principle of this algorithm is to assign the most represented class among the k nearest observations in the data to a new observation, based on a chosen distance.

The only hyperparameter k of this method is the number of nearest observations that we consider to determine the class of a new observation. This parameter allows us to regulate if the comparison of the new observation and the others has to be more or less local.

The Figure 5.1.1 shows a simple example of the k -NN method of classification applied to Euclidean data (in $\mathbb{R}^2$ space) containing two classes. We can see the influence of the choice hyperparameter k of the classification of a new observation. Here, if we choose $k = 1$, i.e. if we look only at the closest observation class, then the new observation will be classified in the class 1. However, if we choose $k = 3$, the most represented class among the 3 closest observations is the class 2, then the new observation will be classified in the class 2.

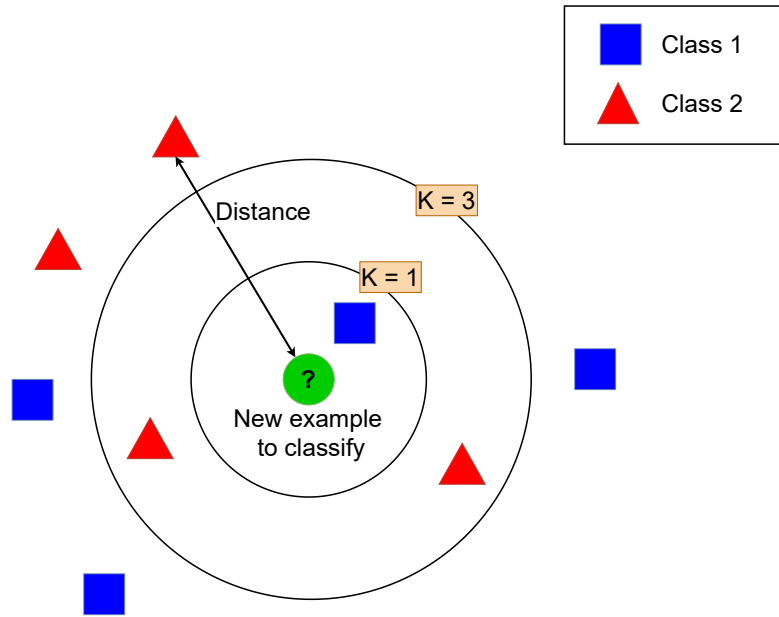


Figure 5.1.1: Influence of the hyperparameter k of the k -Nearest Neighbors method. Inspired from [43]

The k -NN algorithm is a commonly used method for classification problems because it requires metric space. In other words, it can be used for any type of data that has a defined distance. For the classification of time series in the frequency domain, the steps of the k -NN algorithm are as follows:

Steps of the k -Nearest Neighbors Algorithm:

1. **Compute Distances:** For a given new observation $\mathbf{Y}$ with its estimated spectral density $\hat{f}_{\mathbf{Y}}$, compute the distance between $\hat{f}_{\mathbf{Y}}$ and the spectral densities of all N training observations $\hat{f}_{\mathbf{X}_i}$ using the chosen distance metric d .
2. **Identify Nearest Neighbors:** Identify the k observations in the training set that have the smallest distances to $\hat{f}_{\mathbf{Y}}$.
3. **Vote on Class:** Assign the class to $\mathbf{Y}$ based on the majority class among the k nearest neighbors.

4. **Class Assignment:** The new observation $\mathbf{Y}$ is classified into the class that received the most votes among its k nearest neighbors.

The choice of k affects the classifier's sensitivity to local versus global patterns in the data. A smaller k makes the algorithm more sensitive to local variations, while a larger k provides a smoother decision boundary, incorporating more global information.

In literature, we often refer to the training and testing phases of a classifier. The training phase involves the classifier learning from the training data, while the testing phase involves predicting the class of new observations from the testing data. For the k -NN method, there is no traditional training phase as there are no parameters to adjust. Instead, the algorithm relies on the stored training data to make classifications. To enhance understanding, especially in the context of time series data, we can conceptualize the training and testing phases for k -NN as depicted in Figure 5.1.2. During the "training" phase, we store the multitaper spectral estimators of the training observations. In the "testing" phase, we compute the multitaper spectral estimator of the new observation, measure the distances to all training observations, identify the k nearest neighbors, and assign the majority class to the new observation.

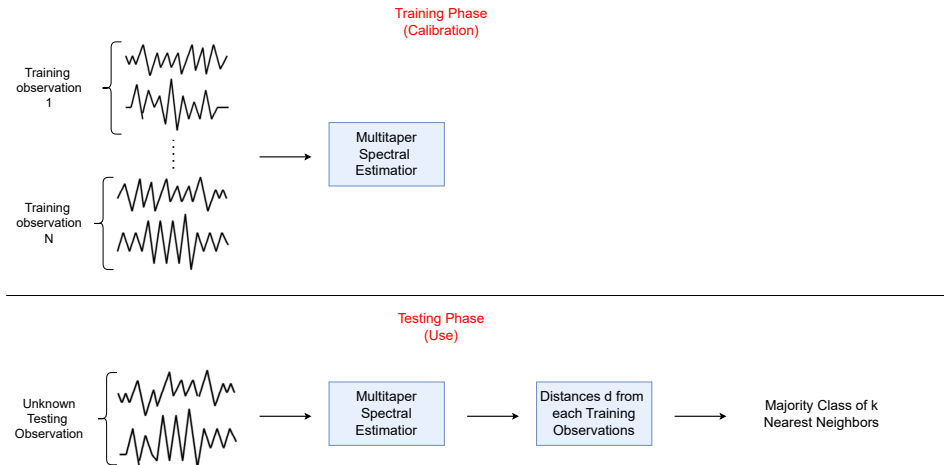


Figure 5.1.2: Conceptual visualization of the training and testing phases of the k -NN method applied to time series data.

In practice, the optimal choice of k hyperparameter is determined using the K -fold cross-validation method, which is described in details in section 7.4.1. In summary, this method involves dividing the data into K subsets, training the model on $K - 1$ subsets, and validating it on the remaining subset. This process is repeated K times, with each subset used exactly once as validation data. The average performance over all K iterations is used to select the optimal k hyperparameter, balancing the trade-off between bias and variance to achieve the best classification performance.

5.2 Minimum Distance to Means

The second algorithm is the Minimum Distance to Means (MDM). In this method, we begin by creating one representative for each class, which is the Karcher mean of the spectral densities of each class with a chosen distance d . Then, we assign the class of the closest representative to the new observation.

In our case, we aim to classify the time series based on their estimated spectra. However, we do not know the true spectral density of each group, denoted ϕ_l . Therefore, the representative of each class $l = 1, \dots, L$ is defined by the Karcher mean with the distance d (see Definition 4.2.3) of the estimated spectral densities of the class:

$$\hat{\phi}_l(\omega) = G_d \left(\{ \hat{f}_{\mathbf{x}_i}(\omega) \}_{\mathbf{x}_i \in \mathcal{C}_l} \right). \quad (5.2.1)$$

The predicted class of the new observation Y , denoted c_Y , with the estimated spectral density $\hat{f}_Y$ will be:

$$c_Y := \arg \min_{l=1, \dots, L} d(\hat{f}_Y, \hat{\phi}_l),$$

where d is the chosen distance metric.

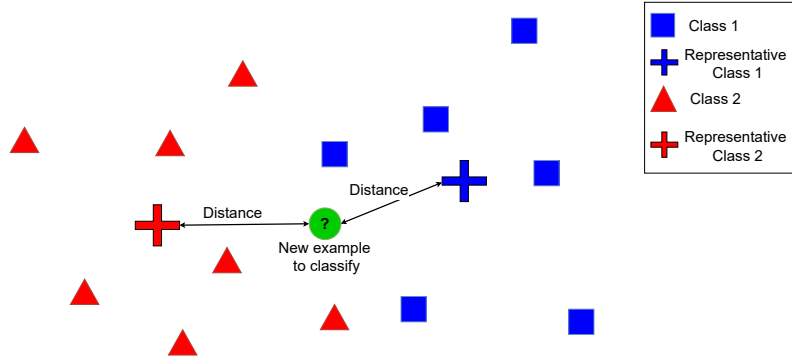


Figure 5.2.1: Visualization of MDM.

The MDM algorithm is particularly useful for classification problems where the Karcher mean spectral density of each class can be a robust representative of that class. The visualization in Figure 5.2.1 illustrates a basic example of the k -NN method of classification applied to Euclidean data (in $\mathbb{R}^2$ space) containing two classes. Each class is represented by its Karcher mean, which is the average mean as we are in a Euclidean space, and a new observation is classified based on its distance to these representatives of the classes.

Steps of the Minimum Distance to Means Algorithm:

1. **Compute Class Representatives :** Calculate the representatives for each class $\hat{\phi}_l(\omega)$ by calculating the Karcher mean spectral density with the chosen distance metric d as defined in the equation 5.2.1.

2. **Compute Distances:** For a given new observation $\mathbf{Y}$ with its estimated spectral density $\hat{f}_{\mathbf{Y}}$, compute the distance between $\hat{f}_{\mathbf{Y}}$ and the Karcher mean spectral density of each class $\hat{\phi}_l$.
3. **Assign Class:** Assign the class to $\mathbf{Y}$ that corresponds to the minimum distance.

The training phase of the MDM method involves calculating the Karcher mean spectral densities based on the chosen distance d for each class from the training data. This step is crucial as it determines the representatives that will be used in the classification of new observations. The testing phase involves estimating the spectral density of a new observation and computing its distances d to the class representatives. The new observation is then assigned to the class whose Karcher mean spectral density is closest to its estimated spectral density.

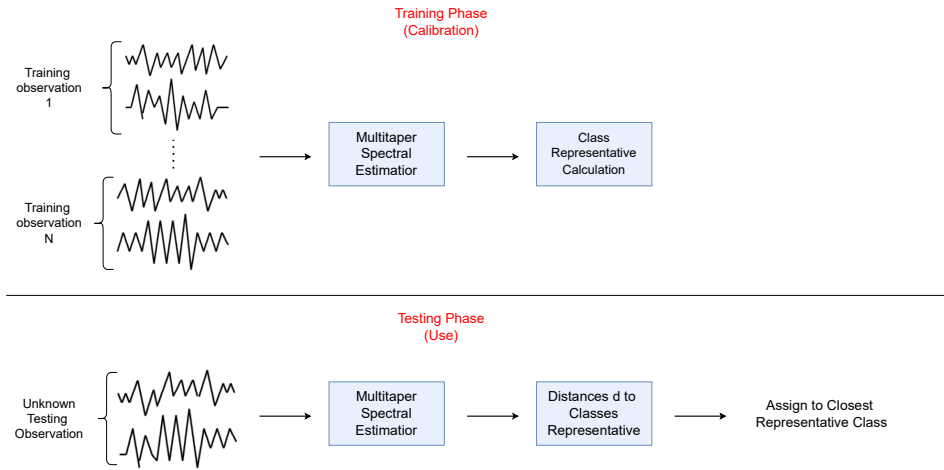


Figure 5.2.2: Conceptual visualization of the training and testing phases of the MDM method applied to time series data. During the training phase, the multitaper spectral estimators of the training observations are used to compute the class representatives. In the testing phase, the multitaper spectral estimator of the new observation is computed, distances to class representatives are measured, and the new observation is assigned to the closest representative class.

The visualization in Figure 5.2.2 illustrates the conceptual workflow of the MDM method. During the training phase, we store the multitaper spectral estimators of the training observations and compute the class representatives. During the testing phase, we compute the multitaper spectral estimator of the new observation, measure the distances to the class representatives, and assign the new observation to the class with the closest representative.

5.3 Sensitivity to Outliers

A notable limitation of the MDM method is its inherent assumption that the distribution of observations within each class is Gaussian or unimodal. This assumption can reduce the flexibility of MDM in handling clusters with non-Gaussian or multimodal distributions. When the data contains outliers, this method can be particularly vulnerable because the Karcher mean, like any mean, is sensitive to extreme values. However, it is important to note that the sensitivity of the MDM method to outliers can be regulated by the choice of distance d used in computing the Karcher mean. The specific distance metric chosen can significantly influence how much outliers affect the mean of the classes.

To illustrate this, consider the example depicted in Figure 5.3.1, where two classes are shown: Class 1 (blue) and Class 2 (red). Class 1 contains a few outliers, which skew the mean. The new observation (shown in green) clearly belongs to Class 1 based on the majority of observations. However, due to the influence of outliers, the Karcher mean of Class 1 is pulled away from its central mass, causing the MDM method to incorrectly classify the new observation as belonging to Class 2.

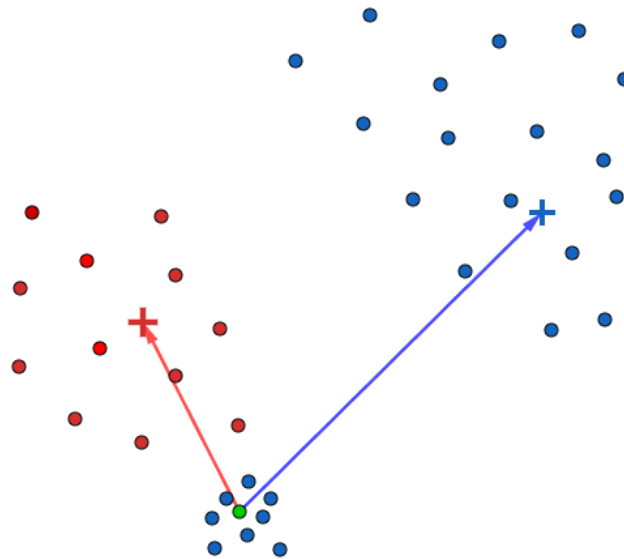


Figure 5.3.1: Example illustrating the sensitivity of MDM to outliers. Class 1 (blue) and Class 2 (red) distributions, with mean values represented by crosses. The green point denotes the new observation, which would be misclassified by the MDM method due to outliers in Class 1.

This example underscores the importance of considering the distributional properties of

the data when choosing between k -NN and MDM methods. While MDM benefits from a stabilized distance metric, its reliance on mean values makes it more susceptible to outliers. However, as we will explore in more details in the next chapter, the choice of distance d associated with the Karcher mean offers a way to regulate this sensitivity, potentially enhancing the robustness of MDM in the presence of outliers. In contrast, k -NN's local comparison approach allows it to mitigate the impact of outliers by focusing on the closest neighbors, offering a more robust classification in such scenarios.

Chapter 6

Spectral Distances

6.1 Introduction to Spectral Distances

In the context of time series classification, the ability to effectively compare different time series is crucial. As discussed in Chapter 5, we are utilizing methods such as k -Nearest Neighbors (k -NN) and Minimum Distance to Means (MDM), both of which rely on the notion of distance to classify time series based on their spectral densities. To perform this comparison in the frequency domain, we require robust and appropriate distance metrics tailored to the spectral density matrices of stationary time series.

Spectral density functions play a vital role in distinguishing between different time series. In signal classification, for instance, we often work with a library of known feature spectral densities, each categorized into specific classes. When presented with an unknown spectral density, we compare it to the feature spectral densities in the library. The similarity between the unknown spectral density and the feature classes is determined by measuring the pairwise distances between them. The closer the distance, the greater the similarity, and the time series is then classified according to its highest similarity to a particular class.

Given the centrality of this comparison process in classification, selecting the right distance metric is paramount. The literature provides a wide array of distance metrics for comparing spectral densities. For instance, [5] offers a comprehensive comparison of several distances in a univariate classification context.

This chapter presents four distance metrics that are commonly used to compare spectral density matrices. Each metric is suited to different characteristics of the data and the manifold on which the spectral densities lie. We begin with the Euclidean distance, a straightforward and widely-used metric for comparing vectors and matrices. Next, we introduce the Log-Euclidean distance, which offers certain advantages in handling positive definite matrices. The third metric is the Affine-Invariant Riemannian (AIR) distance, which leverages the geometric properties of the manifold formed by spectral density matrices. Finally, we explore the Whittle distance, a probability distance that, while not a geometric distance in the strict sense, has important theoretical properties, including consistency, which will be demonstrated in this chapter.

By understanding and applying these distance metrics, we enhance our ability to accurately classify time series based on their spectral properties. The choice of distance metric significantly influences the performance of the classification methods discussed in Chapter 5, and as such, this chapter lays the groundwork for the practical application of these methods in spectral analysis.

6.2 Mathematical Framework for Spectral Distance Comparison

As a reminder, let us assume we take two time series $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t \in \mathbb{N}}$ of length T and dimension P and f_X, f_Y their respective spectral densities. A spectral density is a function $f : [0, \pi] \rightarrow \mathcal{H}_P^+ \subseteq \mathbb{C}^{P \times P} : \omega \mapsto f(\omega)$. For all distances analyzed in this thesis, we will compare the spectral densities by comparing their values (i.e. the spectral matrices) on a chosen subset of frequencies $\Omega \subseteq [0, \pi]$. More formally, all distances d will have the following form in both continuous and discrete cases:

$$d(f_X, f_Y) = \int_{\Omega} \delta(f_X(\omega), f_Y(\omega)) d\omega \approx \sum_{\omega_i \in \Omega} \delta(f_X(\omega_i), f_Y(\omega_i)) \Delta\omega_i, \quad (6.2.1)$$

where δ denotes a distance between two $P \times P$ complex matrices.

We remind the reader that the spectral density function can be interpreted as a curve on the Hermitian Positive Definite (HPD) manifold $\mathcal{H}_P^+$. In this context, the distance between two spectral densities is essentially the comparison of these curves within the manifold, emphasizing the importance of choosing an appropriate distance metric that respects the underlying geometry of the spectral density matrices.

For the purpose of this thesis, we assume equal frequency increments are used, i.e., $\Delta\omega_i$ is a constant. We can then define the distance between the two given HPD curves as:

$$d(f_X, f_Y) = \sum_{\omega_i \in \Omega} \delta(f_X(\omega_i), f_Y(\omega_i)). \quad (6.2.2)$$

Then, finding the right distance d to compare two spectral densities boils down to find which distances δ are the most appropriate to compare both the spectral matrices.

Figure 6.2.1 illustrates the concept of comparing two spectral densities, f_X and f_Y , by visualizing them as curves on the Hermitian Positive Definite (HPD) manifold $\mathcal{H}_P^+$. The blue dashed lines represent the pointwise distances $\delta(f_X(\omega_i), f_Y(\omega_i))$ between the spectral densities at specific frequencies ω_i . The total distance $d(f_X, f_Y)$ is computed by summing these individual distances over the chosen frequency subset $\Omega = [-\pi, \pi]$, using the Euclidean distance as the metric in this example. This visualization underscores how spectral density comparison is performed within the HPD manifold framework.

In a concrete case, we have not the real spectral density. Therefore, we need to estimate it. In Chapter 3.2, we developed two types of estimators: *Periodogram* and *Multitaper Spectral Estimator*. The periodogram presented some limitations, such as the fact that it is

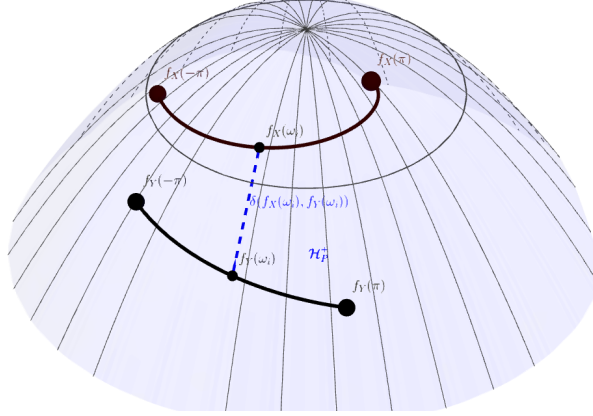


Figure 6.2.1: Two spectral densities f_X and f_Y as curves on the HPD manifold $\mathcal{H}_P^+$. The blue dashed lines represent the pointwise distances $\delta(f_X(\omega_i), f_Y(\omega_i))$ at different frequencies ω_i . The total distance $d(f_X, f_Y)$ is the sum of these distances across all frequencies. The Figure shows an intuitive visualization, i.e., that is not the real manifold.

not consistent, but it can be made consistent by smoothing it with an adequate frequency window.

The primary goal is not always to obtain precise estimates but rather to compare time series using the periodogram. In this context, the immediate usefulness of smoothing may not be evident. When comparing periodogram ordinates at a given set of frequencies, various differences between each spectral density and the corresponding periodogram ordinate are aggregated, potentially averaging out the noise. This means that the comparison inherently involves grouped estimates.

However, both simulation and theoretical results suggest that proper smoothing is highly beneficial [6]. The early work by Coates and Diggle [9] and Diggle and Fisher [12] discussed the usefulness of smoothing. Smoothing enhances the periodogram by reducing variance and highlighting the underlying structure of the spectral density, leading to more reliable comparisons and better classification outcomes [6].

6.3 Euclidean Distance

The Euclidean distance or L_2 distance between two spectral density matrices is the distance induced by Euclidean space. If we consider Euclidean space of all $P \times P$ complex matrices $\mathbb{C}^{P \times P}$, then the natural distance between two matrices $A, B \in \mathbb{C}^{P \times P}$ is the Frobenius distance defined by:

$$\delta_E^2(A, B) := \|A - B\|_F^2 = \sum_{i,j=1}^P \|A_{ij} - B_{ij}\|_2^2, \quad (6.3.1)$$

where $\|\cdot\|_F^2$ and $\|\cdot\|_2$ are the Frobenius norm and the complex norm, respectively.

We then define the Euclidean distance between two spectral densities as follows:

Definition 6.3.1 (Euclidean distance). *Let $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t \in \mathbb{N}}$ of length T be two time series and f_X, f_Y their respective spectral densities. The Euclidean distance between their spectral densities is defined as the sum of the distances between the spectral matrices across the frequency set $\Omega \subseteq [0, \pi]$, i.e.,*

$$d_E^2(f_X, f_Y) := \sum_{\omega_i \in \Omega} \delta_E(f_X(\omega_i), f_Y(\omega_i)). \quad (6.3.2)$$

Although the Frobenius distance seems to be quite natural for comparing complex matrices, it is not aligned with the intrinsic geometry of the space where the spectral matrices lie. Indeed, in Section 4.2, we highlight the fact that spectral matrices belong to the HPD manifold included in the space of complex matrices and have another induced distance proper to this manifold.

Furthermore, we can have a problem when we want to work with the Karcher mean of spectral density points using this distance. Indeed, the Euclidean mean, which is evidently the arithmetic mean, may cause the swelling effect as mentioned in Section 4.2.1. Moreover, the arithmetic mean may not yield a valid spectral density because the HPD manifold is not Euclidean.

We encounter significant issues when attempting to work with the Karcher mean of spectral density matrices using the Euclidean distance. The Euclidean mean, which corresponds to the arithmetic mean, can induce what is known as the swelling effect, as discussed in Section 4.2.1. This effect leads to the mean spectral density matrix becoming unstable, where the determinant of the mean can take extreme values that do not accurately represent the underlying data. Additionally, because the set of Hermitian positive-definite (HPD) matrices forms a curved manifold, rather than a flat Euclidean space, the arithmetic mean of these matrices may not remain within the HPD manifold. As a result, the mean may no longer be a valid spectral density matrix, which can lead to erroneous conclusions or misrepresentations of the spectral characteristics of the time series data.

6.4 Affine-Invariant Riemannian (AIR) Distance

In this section, we leverage the geometric structure of the space of Hermitian positive-definite (HPD) matrices manifold $\mathcal{H}_P^+$, treating it as a Riemannian manifold equipped with the affine-invariant metric. The affine-invariant Riemannian metric is crucial in estimation problems involving Hermitian positive-definite matrices. Indeed, it avoids the swelling effect observed with the Euclidean metric [8], where interpolating between two HPD matrices can result in a matrix with a determinant larger than either of the original matrices, potentially leading to computational instability [35].

We first define the matrix exponential and logarithm:

Definition 6.4.1 (Matrix exponential). *Let $A \in \mathbb{C}^{P \times P}$ be a complex matrix. The matrix exponential of A is defined as the unique matrix e^A such that*

$$e^A := \sum_{n=0}^{\infty} \frac{A^n}{n!}. \quad (6.4.1)$$

Definition 6.4.2 (Matrix logarithm). *Let $A \in \mathbb{C}^{P \times P}$ be an complex matrix. The matrix logarithm of A is defined as the matrix $\log(A)$ such that $e^{\log(A)} = A$.*

Since the exponential function is not bijective for complex numbers (for instance, $e^{\pi i} = e^{3\pi i} = -1$), numbers can possess multiple complex logarithms. Consequently, some matrices can also have more than one logarithm. A complex matrix has a logarithm if and only if it is invertible [15]. Although the logarithm is generally not unique, if a matrix does not have negative real eigenvalues, there exists a unique logarithm whose eigenvalues are all situated within the strip $z \in \mathbb{C} \mid -\pi < \text{Im } z < \pi$. This unique logarithm is referred to as the principal logarithm. In this thesis, we work with spectral matrices which are positive-definite (cf 3), therefore we have unique logarithm.

The Affine-Invariant Riemannian distance is derived from the geometry of the HPD manifold $\mathcal{H}_P^+$ [37]. This distance between two points A and B in the manifold $\mathcal{H}_P^+$ is defined as the geodesic distance:

$$\delta_R^2(A, B) := \|\log(A^{-1/2}BA^{-1/2})\|_{\mathcal{F}}^2.$$

Thus, the Affine-Invariant Riemannian distance for spectral densities is formally defined as follows:

Definition 6.4.3 (Affine-Invariant Riemannian (AIR) distance). *Let $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t \in \mathbb{N}}$ be two time series of length T with respective spectral densities f_X and f_Y . The Affine-Invariant Riemannian distance between their spectral densities, across the frequency set $\Omega \subseteq [0, \pi]$, is given by:*

$$d_{AIR}(f_X, f_Y) := \sum_{\omega_i \in \Omega} \delta_R(f_X(\omega_i), f_Y(\omega_i)). \quad (6.4.2)$$

The definition of the Affine-Invariant Riemannian distance as the length of geodesic ensures that it has all the properties of a distance function. It has other interesting properties, one of them is called the congruence invariance and is defined by [36]

Property 6.4.4 (Congruence invariance). *Let $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t \in \mathbb{N}}$ be two time series of length T with respective spectral densities f_X and f_Y . Let A be an invertible matrix. The mapping $f \mapsto A * f$, with $(A * f)(\omega) := Af(\omega)A^T, \forall \omega \in [-\pi, \pi]$ is called a congruence. Then, the AIR distance is invariant under any congruence, i.e.,*

$$d_{AIR}(f_X, f_Y) = d_{AIR}(A * f_X, A * f_Y). \quad (6.4.3)$$

The Affine-Invariant Riemannian distance possesses several beneficial properties and enhances the estimation of the Karcher mean [8, 10]. The Karcher mean eliminates the swelling effect by utilizing the intrinsic geometry of the manifold, resulting in a more accurate and representative average. The determinant is continuous and monotonic along the geodesic connecting two spectral densities, ensuring stability and consistency. The Karcher mean with respect to this distance satisfies all the ALM properties (c.f 4.2.1).

The Affine-Invariant Riemannian distance between spectral density matrices on the HPD manifold is particularly advantageous for classification and comparison tasks. By exploiting the intrinsic geometry of the HPD manifold, this Riemannian distance provides a robust and theoretically sound method for comparing spectral densities, leading to more accurate and meaningful classifications of time series data.

6.5 Log-Euclidean Distance

The Log-Euclidean distance between spectral density matrices uses the matrix logarithm to map HPD matrices into a vector space where Euclidean geometry applies. This approach simplifies computations while preserving many desirable properties of the Riemannian distance.

Let A and B be two Hermitian positive-definite (HPD) matrices. The Log-Euclidean distance between A and B is defined as:

$$\delta_{LE}^2(A, B) := \|\log(A) - \log(B)\|_F, \quad (6.5.1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The Log-Euclidean distance between two spectral densities is defined analogously to the Euclidean distance but using the Log-Euclidean metric:

Definition 6.5.1 (Log-Euclidean distance). *Let $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t \in \mathbb{N}}$ be two time series with respective spectral densities f_X and f_Y . The Log-Euclidean distance between these spectral densities, across the frequency set $\Omega \subseteq [0, \pi]$, is given by:*

$$d_{LE}(f_X, f_Y) := \sum_{\omega_i \in \Omega} \delta_{LE}(f_X(\omega_i), f_Y(\omega_i)). \quad (6.5.2)$$

The Log-Euclidean distance provides a balance between computational simplicity and geometric fidelity. By working in the logarithmic space, it mitigates some limitations of the Euclidean distance while avoiding the computational complexities of the full Riemannian framework.

6.6 Whittle Distance

The Whittle distance, is fundamentally different from the previously discussed distances as it is based on the likelihood function of a time series. It is based on the Whittle likelihood,

which is a frequency domain approximation to the exact likelihood, introduced in 1953 by mathematician and statistician Peter Whittle [48].

Although the likelihood method is often used as a standard method in parametric spectral analysis [44], it can also be applied for classification. The idea is to measure how likely a newly observed time series belongs to a given class. More formally, assuming that we have L classes $\mathcal{C}_l$, $l = 1, \dots, L$ with respective spectral densities ϕ_l , we can compute the likelihood of an observed time series X assuming it belongs to class l . Here, we use the Whittle approximation of this log-likelihood in the frequency-domain approximation, also named Whittle likelihood in the literature and defined as [40, p. 431]:

$$\log p_l(X) := \sum_{0 < \omega_i < \pi} \left[-\log |\phi_l(\omega_i)| - \text{tr}\{I_X(\omega_i)\phi_l^{-1}(\omega_i)\} \right] \quad (6.6.1)$$

where I_X is the periodogram of the time series X , $|A|$ is the determinant of A , and tr is the trace operator.

We compute this for each class and assign the observation X to the class with the highest likelihood. We can observe that this method is equivalent in its reasoning to the MDM method. The goal is to compare a new observation with each representative (here the mean spectral density) of the class, and assign this new observation to the same class as the closer representative. Indeed, if we apply the MDM method with the Whittle likelihood as distance, we have the same method that the one described in this section. The Whittle likelihood provides a measure of similarity between two spectral densities. However, since a higher likelihood indicates closer similarity, which is the opposite of classical distance measures, we define the Whittle distance as the negative of the Whittle likelihood:

Definition 6.6.1 (Whittle Distance). *Let X and Y be two time series of length T with respective spectral densities f_X and f_Y . The Whittle distance between their spectral densities, across the frequency set $\Omega \subseteq [0, \pi]$, is given by:*

$$d_W(f_X, f_Y) := \sum_{\omega_i \in \Omega} \left[\log |f_Y(\omega_i)| + \text{tr}\{f_X(\omega_i)f_Y^{-1}(\omega_i)\} \right] \quad (6.6.2)$$

Although it is effective in spectral analysis and time series classification, it does not strictly conform to the geometric definition of a distance. Indeed, the Whittle distance does not verify the distance properties. The most important aspect to notice is the asymmetry aspect, i.e. in general

$$d_W(f_X, f_Y) \neq d_W(f_Y, f_X) \quad (6.6.3)$$

for two distinct spectral densities f_X, f_Y .

The Whittle likelihood or Whittle distance is usually used with the method described above, i.e. with the MDM method (cf. Chapter 5). However, as we also describe it as a pseudo-distance, we can easily apply the k -NN (cf. Chapter 5) method too by using the Whittle distance as the distance to compute the k closest points.

As mentioned before, the Whittle distance is not symmetric. Consequently, the choice of the input order of the distance is important, especially in a concrete case where the

true spectral densities are unknown, and therefore estimated. We can see from the Whittle distance definition that the determinant and the inverse of the second input are used. However, if our spectral density estimator is not stable, then the determinant and the inverse can explode. That is why we have to be careful with this potential problem. Addressing this issue requires a different approach depending on the classification methods we use.

MDM The MDM method has been the subject of more extensive study, as evidenced by the article [18] or in the book [40]. Firstly, the standard approach is to use the periodogram to estimate the spectral density of the observed time series. Then, the representative of the class l , denoted ϕ_l , is obtained using the Euclidean mean, i.e. the average periodogram of class l . Once more, we use the Euclidean mean, therefore it is also sensitive to the swelling effect as explained in the Euclidean distance section 6.3.

The new time series observed time series Y that we want to classify is then compared to each representative using the order $d_W(I_Y, \phi_l)$ where I_X is the periodogram of Y . Therefore, only representative of classes will appear in the second input, which means that only the representatives of classes needs to be consistent. As the determinant and inverse of ϕ_l , we need an extra step to smooth the representative with a sufficient smoothing parameter, to ensure a positive definite estimator and avoid that the determinant and inverse explode. In conclusion, we work with the periodogram for each observation and with the multitaper spectral estimators with $B > P$ tapering functions for the representative, for the representative.

k -NN The k -NN equipped with the Whittle distance is less commonly used. In this method, we directly compare the observed estimated spectral densities, the two inputs in the Whittle distance are the estimated spectral densities of observations. Therefore, here, all observations will appear in the second distance input. Then, all estimated spectral density of the observation must be definite-positive for the same reasoning as explained for the classes representative in the previous paragraph. In conclusion, we work with the multitaper estimation of the spectral density for each observation.

6.6.1 Whittle Consistency

In this section, our goal will be to prove the consistency of the MDM method equipped with the Whittle distance in a multivariate context. This property was initially proven for the univariate case in [18] and has been extended to the multivariate case in this work.

The consistency of a classification is an important property that ensures the reliability of this method. Intuitively, the consistency of a classification method means that as the size of the training dataset and the length of the time series increase, the method's ability to correctly classify new observations improves, ultimately approaching perfect classification. In other words, with enough data and sufficiently long time series, the probability of misclassification converges to zero.

Suppose we aim to classify multivariate time series $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$ of length T and dimension P into L classes, denoted $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_L$, of size $N_1, N_2, \dots, N_L$, respectively. Let $I_{\mathbf{X}_i}$ be the periodogram of $\mathbf{X}_i$. We approximate the true spectral density of the class i , denoted ϕ_i , by

$$\hat{\phi}_i(\omega) = \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{X}_k \in \mathcal{C}_i} \hat{f}_{\mathbf{X}_k}(\omega), \quad (6.6.4)$$

where $\hat{f}_{\mathbf{X}_k}(\cdot)$ is the multitaper spectral estimator of the time series $\mathbf{X}_k$ with $B \geq P$ orthogonal tapering functions, ensuring the estimator is positive definite. The choice of the tapering functions are the DPSS and a given time-bandwidth parameter n_w .

Let Y be a new observed time series with periodogram I . We define the discriminant statistic by

$$\mathcal{D}_W(\phi_1, \phi_2; I) = \frac{1}{T} \left(d_W(I, \phi_1) - d_W(I, \phi_2) \right), \quad (6.6.5)$$

We classify the new time series Y into class $\mathcal{C}_1$ if $\mathcal{D}_W(\hat{\phi}_1, \hat{\phi}_2; I) \leq 0$, and into class $\mathcal{C}_2$ if $\mathcal{D}_W(\hat{\phi}_1, \hat{\phi}_2; I) > 0$.

Our goal is to show that the misclassification probability converges to zero as the time series length and the class sizes tend to infinity. More formally, we state the following theorem:

Theorem 6.6.2 (Whittle Consistency). *Let $\mathbf{X}$ be a time series of length T and dimension P , and assume the true class of $\mathbf{X}$ is $\mathcal{C}_l$. Suppose $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$ are multivariate time series as described above, classified into L classes $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_L$. Assume that:*

1. *The true spectral densities of each class $\phi_1, \phi_2, \dots, \phi_L$ are distinct and unknown. They are estimated by*

$$\hat{\phi}_i(\omega) = \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{X}_k \in \mathcal{C}_i} \hat{f}_{\mathbf{X}_k}(\omega), \quad (6.6.6)$$

where $\hat{f}_{\mathbf{X}_k}(\cdot)$ is the multitaper spectral estimator of the time series $\mathbf{X}_k$ with $B \geq P$ orthogonal tapering functions, ensuring the estimator is positive definite. The choice of the tapering functions is the DPSS and a given time-bandwidth parameter n_w .

2. *The periodograms of $\mathbf{X}$ is denoted I .*
3. *The class sizes $N_1, N_2, \dots, N_L$ increase such that $\min(N_1, N_2, \dots, N_L) \rightarrow \infty$ as $N \rightarrow \infty$.*

Then, the misclassification probability converges to zero as the time series length T and the class sizes $N_1, N_2, \dots, N_L$ tend to infinity. Specifically, we have

$$\lim_{N_1, N_2, \dots, N_L \rightarrow \infty} \lim_{T \rightarrow \infty} \left[P \left(\mathcal{D}_W(\hat{\phi}_1, \hat{\phi}_2; I) > 0 \mid \mathbf{X} \in \mathcal{C}_1 \right) \right] = 0. \quad (6.6.7)$$

6.6.1.1 Proof of the Consistency

Proving the consistency of the MDM method with Whittle's distance is equivalent to proving the inequality (6.6.7). To approach this proof, we will first consider the scenario where we know the true spectral densities of the classes, denoted ϕ_l for $l = 1, \dots, L$. We call this version of the Consistency Theorem, the Weak Consistency Theorem. We begin by establishing the following proposition:

Theorem 6.6.3 (Weak Consistency Theorem). *Suppose the true spectral densities of the classes $\phi_1, \phi_2, \dots, \phi_L$ are known. Then, the misclassification probability converges to zero as the time series length T and the class sizes $N_1, \dots, N_L$ tend to infinity. Formally, we have:*

$$\lim_{N_1, \dots, N_L \rightarrow \infty} \lim_{T \rightarrow \infty} \left[P\left(\mathcal{D}_W(\phi_1, \phi_2; I) > 0 \mid Y \in \mathcal{C}_1\right) \right] = 0. \quad (6.6.8)$$

To prove this proposition, we will show that the discriminant statistic $\mathcal{D}_W(\phi_1, \phi_2; I)$ converges in probability to a negative constant.

Once we have established Theorem 6.6.3, we will extend the result to the general case, by proving the consistency theorem (6.6.7) without assuming knowledge of the true spectral densities of the classes. This involves replacing the true spectral densities with their estimates and showing that the estimated discriminant statistic also converges to a negative constant, thereby ensuring the overall consistency of the classification method.

First of all, we can prove that the expectation of the discriminant statistic converges to a negative number.

Lemma 6.6.4 (Non-Positiveness of Expected Discriminant Statistic). *First of all, we will show that*

$$\mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] \leq 0 \quad \text{as } T \rightarrow \infty, \quad (6.6.9)$$

where $\mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = \mathbb{E}[\mathcal{D}_W(\phi_1, \phi_2; I) \mid Y \in \mathcal{C}_1]$.

Proof. Firstly, let us develop the expectation as follows

$$\begin{aligned} \mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] &= \frac{1}{T} \mathbb{E}_1 \left[\sum_{0 < \omega_i < 1/2} \left(\log \left(|\phi_2^{-1}(\omega_i)| \cdot |\phi_1(\omega_i)| \right) \right. \right. \\ &\quad \left. \left. + \text{tr} \left\{ \left(\phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i) \right) I(\omega_i) \right\} \right) \right]. \end{aligned}$$

Here,

$$\mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = \overbrace{\frac{1}{T} \sum_{0 < \omega_i < 1/2} \left(\log \left(|\phi_2^{-1}(\omega_i)| \cdot |\phi_1(\omega_i)| \right) \right)}^{(*)} + \underbrace{\frac{1}{T} \sum_{0 < \omega_i < 1/2} \left(\text{tr} \left\{ \left(\phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i) \right) \mathbb{E}_1(I(\omega_i)) \right\} \right)}_{(**)}.$$

If we develop the second of $\mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)]$, i.e. $(**)$, knowing that $\mathbb{E}_1[I(\omega)] = \phi_1(\omega) + O(T^{-1})$, we have that

$$\begin{aligned} (**) &= \frac{1}{T} \sum_{0 < \omega_i < 1/2} \left(\text{tr} \left\{ \left(\phi_1^{-1}(\omega) - \phi_2^{-1}(\omega) \right) \left(\phi_1(\omega_i) + O(T^{-1}) \right) \right\} \right) \\ &= \frac{1}{T} \sum_{0 < \omega_i < 1/2} \left(P - \text{tr} \left\{ \left(\phi_2^{-1}(\omega_i) \phi_1(\omega) \right) \right\} \right) + O(T^{-1}). \end{aligned}$$

Then, to show that $\mathbb{E}_1[\mathcal{D}_W(\phi_1, \phi_2; I)]$ is non-positive as $T \rightarrow \infty$, it suffices to show that

$$\log \left(|\phi_2^{-1}(\omega_i)| \cdot |\phi_1(\omega_i)| \right) - \text{tr} \left\{ \phi_2^{-1}(\omega_i) \phi_1(\omega_i) \right\} + P \leq 0 \quad \forall \omega_i. \quad (6.6.10)$$

Let us develop this expression:

$$\begin{aligned} \log \left(|\phi_2^{-1}(\omega_i)| \cdot |\phi_1(\omega_i)| \right) - \text{tr} \left\{ \phi_2^{-1}(\omega_i) \phi_1(\omega_i) \right\} + P &= \log \left(|\phi_2^{-1}(\omega_i) \phi_1(\omega_i)| \right) \\ &\quad - \text{tr} \left\{ \phi_2^{-1}(\omega_i) \phi_1(\omega_i) \right\} + P \end{aligned}$$

Let $A := \phi_2^{-1}(\omega_i) \phi_1(\omega_i) \in \mathbb{R}^{P \times P}$ with eigenvalues $(\lambda_k)_{k=1}^P$, then

$$\begin{aligned} (3) &= \log \left(|A| \right) - \text{tr} \left\{ A \right\} + P \\ &= \log \left(\prod_{k=1}^P \lambda_k \right) - \sum_{k=1}^P \lambda_k + P \\ &= \sum_{k=1}^P \log \left(\lambda_k \right) - \sum_{k=1}^P \lambda_k + P \\ &= \sum_{k=1}^P \left(\log \left(\lambda_k \right) - \lambda_k + 1 \right). \end{aligned}$$

Finally, we know that $\log(z) - z + 1 \leq 0$ for all $z \in \mathbb{R}$, which concludes the proof.

Additionally, note that it is equal to zero when all $\lambda_k = 1$, which implies $A = \text{Id}$ and therefore $\phi_1 = \phi_2$. $\square$

Then, we want to prove that the variance of the discriminant statistic converges to zero.

Lemma 6.6.5 (Convergence of the Variance of the Discriminant Statistic). *Let us show that*

$$\mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = 0 \quad \text{as } T \rightarrow \infty \quad (6.6.11)$$

$$\text{where } \mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = \mathbb{V}\left[\mathcal{D}_W(\phi_1, \phi_2; I) \middle| Y \in \mathcal{C}_1\right].$$

Proof. We start by developing the conditional variance:

$$\begin{aligned} \mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] &= \mathbb{V}_1\left[\frac{1}{T} \sum_{0 < \omega_i < 1/2} \left(\log\left(|\phi_2^{-1}(\omega_i)| \cdot |\phi_1(\omega_i)|\right) \right. \right. \\ &\quad \left. \left. + \text{tr}\left\{\left(\phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i)\right)I(\omega_i)\right\}\right) \right] \\ &= \frac{1}{T^2} \mathbb{V}_1\left[\sum_{0 < \omega_i < 1/2} \underbrace{\text{tr}\left\{\left(\phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i)\right)I(\omega_i)\right\}}_{Q_i}\right]. \end{aligned}$$

Let $Q_i := \text{tr}\left\{\left(\phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i)\right)I(\omega_i)\right\}$, then

$$\mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = \frac{1}{T^2} \sum_i \mathbb{V}_1(Q_i) + \frac{1}{T^2} \sum_i \sum_{j \neq i} \text{Cov}_1(Q_i, Q_j). \quad (6.6.12)$$

Let us compute the expression $\mathbb{V}_1(Q_i)$ and $\text{Cov}_1(Q_i, Q_j)$.

Firstly, we can develop the trace into a sum. To facilitate readability, we will let $A_i := \phi_1^{-1}(\omega_i) - \phi_2^{-1}(\omega_i)$, which is not a random variable. Furthermore, we will denote the element (k, l) of the matrix $I(\omega_i)$ as $(I(\omega_i))_{(k, l)} = I_{(k, l)}(\omega_i)$. Therefore, we have

$$\mathbb{V}_1(Q_i) = \mathbb{V}_1\left[\sum_{k, l=1}^P (A_i)_{(k, l)} I_{(l, k)}(\omega_i)\right].$$

By the property 3.2.4, we know that

$$\mathbb{V}\left(I_{(l, k)}(\omega_i)\right) = \left(\phi_1^2(\omega_i)\right)_{(k, l)} + \left(\phi_1(\omega_i)\right)_{(k, k)} \left(\phi_1(\omega_i)\right)_{(l, l)} + O(T^{-1}). \quad (6.6.13)$$

We can then use the Cauchy-Schwarz inequality to show that if $\sup_i \mathbb{V}(\mathbf{X}_i) < \infty$, then $\mathbb{V}\left(\sum_{i=1}^m \mathbf{X}_i\right) \leq m \sum_{i=1}^m \mathbb{V}(\mathbf{X}_i)$. Therefore, the variance of Q_i can be upper bounded by $v_i + O(T^{-1})$ is a constant:

$$\begin{aligned} \mathbb{V}_1(Q_i) &\leq kl \sum_{k, l=1}^P (A_i)_{(k, l)}^2 \mathbb{V}_1\left(I_{(l, k)}(\omega_i)\right) \\ &\leq kl \sum_{k, l=1}^P (A_i)_{(k, l)}^2 \left(\left(\phi_1^2(\omega_i)\right)_{(k, l)} + \left(\phi_1(\omega_i)\right)_{(k, k)} \left(\phi_1(\omega_i)\right)_{(l, l)} + O(T^{-1}) \right) =: v_i + O(T^{-1}). \end{aligned}$$

Therefore,

$$\frac{1}{T^2} \sum_i \mathbb{V}_1(Q_i) \leq \frac{1}{T^2} \sum_i (v_i + O(T^{-1})) \approx O(T^{-1}). \quad (6.6.14)$$

Secondly, we can develop the second term in the same way:

$$\text{Cov}_1(Q_i, Q_j) = \text{Cov}_1\left(\sum_{k_1, l_1=1}^P (A_i)_{(k_1, l_1)} I_{(l_1, k_1)}(\omega_i), \sum_{k_2, l_2=1}^P (A_j)_{(k_2, l_2)} I_{(l_2, k_2)}(\omega_j)\right).$$

By the property 3.2.5, we know that $\forall w_i, w_j \in [-\pi, \pi]$, with $w_i \neq w_j$ and $k_i, l_i = 1, \dots, P$,

$$\text{Cov}_1(I_{(l_1, k_1)}(\omega_i), I_{(l_2, k_2)}(\omega_j)) \approx O(T^{-1}).$$

Hence,

$$\text{Cov}_1(Q_i, Q_j) \approx O(T^{-1}) \quad (6.6.15)$$

Finally, using 6.6.14 and 6.6.15, we can bound 6.6.12 as follows

$$\begin{aligned} \mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] &\approx \frac{1}{T^2} \sum_{w_i} O(T^{-1}) + \frac{1}{T^2} \sum_{w_i} \sum_{\substack{w_j \\ w_j \neq w_i}} O(T^{-1}) \\ &\approx O(T^{-1}) \end{aligned}$$

Hence, we can conclude that

$$\mathbb{V}_1[\mathcal{D}_W(\phi_1, \phi_2; I)] = 0 \quad \text{as } T \rightarrow \infty,$$

which proves the proposition $\square$

Now, we can simply prove the Weak Consistency Theorem with the true spectral densities of the classes known by assembling the lemmas above.

Proof of the Weak Consistency Theorem 6.6.3. By proposition 6.6.4, we know that the conditional expectation of $\mathcal{D}_W(\phi_1, \phi_2; I)$ converges to a negative value as $T \rightarrow \infty$. Furthermore, by the proposition 6.6.5, the conditional variance of $\mathcal{D}_W(\phi_1, \phi_2; I)$ converges to zero as $T \rightarrow \infty$. Therefore, by Chebyshev's inequality [2], we conclude that

$$\lim_{T \rightarrow \infty} \left[P\left(\mathcal{D}_W(\phi_1, \phi_2; I) > 0 \mid Y \in \mathcal{C}_1\right) \right] = 0.$$

$\square$

Finally, we can return to the desired Consistency Theorem

Proof of the Consistency Theorem 6.6.2. By the strong law of large numbers, $\mathcal{D}_W(\hat{\phi}_1, \hat{\phi}_2; I) \xrightarrow{\text{a.s.}} \mathcal{D}_W(\phi_1, \phi_2; I)$ as $N_1, N_2 \rightarrow \infty$, and the result follows immediately from the previous section. $\square$

6.7 Comparative Analysis of Karcher Means Associated with Different Spectral Distances

This section builds on the discussion in Section 5.3 regarding the sensitivity of the MDM method to outliers. In the context of the Minimal Distance to Mean (MDM) method, the Karcher mean serves as the representative element of a set of estimated spectral densities for each class, based on a given distance metric d . This section delves into an empirical comparison of the robustness of the Karcher means when computed with different spectral distance metrics: Euclidean, AIR, and log-Euclidean.

To investigate this, we simulate 100 time series, introducing 10% outliers to assess the sensitivity of the Karcher mean under different distances. Specifically, 90 time series are generated from a specified VAR(1) process, while the remaining 10 are outliers generated from a VAR(1) process with distinct parameters. The precise details of the VAR(p) processes will be elaborated in the following chapter, Simulation Study 7.

The Karcher means of these 100 time series are then computed using the Euclidean, AIR, and log-Euclidean distances. The objective is to observe the extent to which these means are influenced by the presence of outliers.

Figure 6.7.1 provides a visual representation of how the different Karcher means respond to the outliers. The Euclidean distance proves to be more sensitive to outlier effects, particularly in the high-frequency range, where peaks are more pronounced.

However, the AIR and log-Euclidean distances exhibit a more pronounced bias at lower frequencies. This bias is attributed to the intrinsic properties of these distance metrics, where the bias tends to be multiplicative or proportional to the spectral density's value. Notably, Joris Chau addresses this issue in his thesis, proposing a multiplicative correction term in Theorem 3.4.1 [8]. However, this correction is beyond the scope of this thesis.

The findings from this empirical analysis underscore the robustness of the AIR and log-Euclidean distances in producing Karcher means that are less susceptible to outlier influence. Such robustness is particularly advantageous when dealing with real-world data, which often includes outliers due to various factors such as human error or equipment malfunction. In applications where robustness against outliers is a critical consideration, this study advocates the use of AIR and log-Euclidean distances.

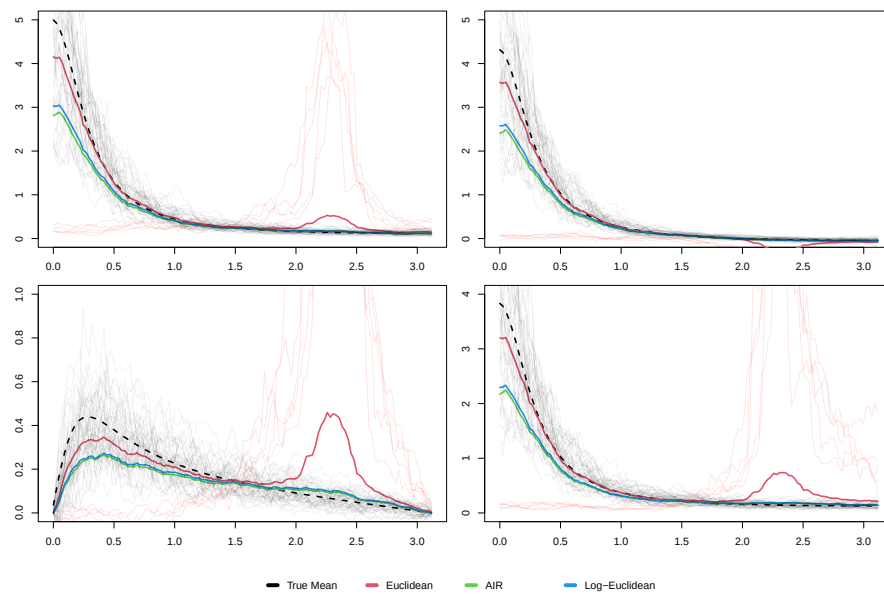


Figure 6.7.1: Comparison of Karcher means computed with different distances. The transparent black and red lines represent the 90 observations from the primary VAR(1) process and the 10 outliers from a secondary VAR(1) process, respectively. The dashed black line indicates the true spectral density of the first process. The red, green, and blue lines depict the Karcher means computed using the Euclidean, AIR, and log-Euclidean distances, respectively.

Chapter 7

Application : Simulation Study

In this chapter, we apply the methods developed in this thesis, namely Minimum Distance to Means (MDM) and k -Nearest Neighbors (k -NN), using different distance metrics on simulated data. The purpose of this simulation study is to create a controlled environment where we can evaluate the performance of our classification algorithms. By using simulated data, we can ensure that the underlying assumptions of our models are satisfied, which helps in assessing the true capabilities of our methods. Additionally, the use of simulated data allows us to explore the effectiveness of frequency domain analysis in classifying time series that share similar covariance structures but differ in their autoregressive parameters.

7.1 Vector Autoregression (VAR)

Vector autoregression (VAR) is a statistical technique designed to capture the dynamic relationship between multiple variables over time. It extends the univariate autoregressive model to accommodate multivariate time series, making it a type of stochastic process model. VAR models are frequently employed in macroeconomic econometrics [41] and the natural sciences[47].

In a VAR model, each variable's evolution over time is described by an equation that includes its own past values, the past values of other variables in the model, and an error term. Unlike structural models with simultaneous equations, VAR models require less prior knowledge about the underlying forces affecting the variables. The primary requirement is identifying a set of variables that are hypothesized to influence each other over time.

A VAR model tracks the progression of k endogenous variables over time, with each time period labeled $t = 1, \dots, T$. These variables are represented in a vector $\mathbf{X}(t)$, which has a length of k .

The order of a VAR model, denoted p , indicates the number of past time periods included in the model. A p -th order VAR model incorporates lags from the last p periods and is formulated as:

$$\mathbf{X}(t) = \mathbf{c} + \Phi_1 \mathbf{X}(t-1) + \Phi_2 \mathbf{X}(t-2) + \dots + \Phi_p \mathbf{X}(t-p) + \mathbf{Z}_t \quad (7.1.1)$$

where:

- $\mathbf{X}(t)$ is the vector of endogenous variables.
- $\mathbf{c}$ is a k -vector of constants (intercepts).
- Φ_i are time-invariant $k \times k$ matrices.
- $\mathbf{Z}_t$ is a k -vector of error terms, which must satisfy the following conditions:
 - $\mathbb{E}(\mathbf{Z}_t) = 0$ (each error term has a mean of zero),
 - $\mathbb{E}(\mathbf{Z}_t \mathbf{Z}_t^T) = \mathbf{\Omega}$ (the contemporaneous covariance matrix of error terms is a $k \times k$ positive-semidefinite matrix),
 - $\mathbb{E}(\mathbf{Z}_t \mathbf{Z}_{t-k}^T) = 0$ for any non-zero k (no serial correlation in individual error terms).

Figures 2.0.2, 2.1.1 and 3.1.1 illustrate an example of a VAR(2) process and its characteristics.

Stationarity A VAR(p) process such as defined in the equation 7.1.1 is stationary if the roots of the polynomial $\det(I - \Phi_1 z - \dots - \Phi_p z^p)$ lie inside the unit circle for all $|z| \leq 1$. This condition ensures that the time series does not exhibit explosive behavior and that its statistical properties remain constant over time. By default, it includes a burn-in period of 100 observations to ensure the stationarity of the generated series.

7.2 Data Generation

To generate artificial data, we utilize the ‘rArma()’ function from the ‘pdSpecEst’ package[7]. This function is designed to simulate time series data with a predefined autoregressive moving average structure.

Here we will simulate bidimensional time series data using VAR(1) processes, which means that the value of the time series at time t is proportional to its value at time $t - 1$ with an error term. Furthermore, we consider the mean of the process equal to zero without losing generality. More formally, let $\mathbf{X}(t) = (\mathbf{X}_{1,t}, \mathbf{X}_{2,t})$ be a bivariate time series following a VAR(1) process, represented as:

$$\mathbf{X}(t) = \Phi \mathbf{X}(t-1) + \mathbf{Z}_t \quad (7.2.1)$$

where $\Phi \in \mathbb{R}^{2 \times 2}$ and $\mathbf{Z}_t \sim \mathcal{N}_2(0, \Sigma)$. Here, Φ is the autoregressive coefficient matrix, and $\mathbf{Z}_t$ is a vector of white noise with covariance matrix Σ .

We generate three groups of time series, each with different parameters for Φ but sharing the same covariance matrix. This approach highlights the utility of the frequency domain for classification, as it allows us to differentiate between time series based on their spectral properties rather than their covariance structures alone.

The covariance of a VAR(1) process $\mathbf{X}$ of dimension P , is denoted $\Gamma_X := \text{Cov}(\mathbf{X}(t))$. We can show the following property:

Property 7.2.1. Let $\mathbf{X}$ be a VAR(1) process of dimension P defined by the model :

$$\mathbf{X}(t) = \Phi \mathbf{X}(t-1) + \mathbf{Z}_t, \quad (7.2.2)$$

where $\mathbf{Z}_t \sim \mathcal{N}_P(0, \Sigma)$.

The covariance matrix of $\mathbf{X}$, denoted $\Gamma_X := \text{Cov}(\mathbf{X}(t))$:

$$\text{vec}(\Gamma_X) = \left(\text{Id}_{P^2 \times P^2} - \Phi \otimes \Phi \right)^{-1} \text{vec}(\Sigma), \quad (7.2.3)$$

where $\text{vec}(\cdot)$ is the vectorization, $\text{Id}_{P^2 \times P^2}$ is the $P^2 \times P^2$ identity matrix and $\otimes$ is the Kronecker product [32].

Proof. If we take the variance in both sides of the equation 7.2.2, we obtain

$$\mathbb{V}(\mathbf{X}(t)) = \Phi \mathbb{V}(\mathbf{X}(t-1)) \Phi^T + \Sigma. \quad (7.2.4)$$

By stationarity, the covariance matrix does not vary over time, i.e. $\mathbb{V}(\mathbf{X}(t)) = \mathbb{V}(\mathbf{X}(t-1)) = \Gamma_X$. So we have

$$\Gamma_X = \Phi \Gamma_X \Phi^T + \Sigma. \quad (7.2.5)$$

Then, if we apply the vectorization in both sides and use the Roth's relationship [32] that implies $\text{vec}(\Phi \Gamma_X \Phi^T) = (\Phi \otimes \Phi) \text{vec}(\Gamma_X)$, with $\otimes$ the Kronecker product, the equation above can be rewritten as

$$\text{vec}(\Gamma_X) = (\Phi \otimes \Phi) \text{vec}(\Gamma_X) + \text{vec}(\Sigma). \quad (7.2.6)$$

Finally, by isolating the covariance matrix, we prove the property

$$\text{vec}(\Gamma_X) = \left(\text{Id}_{P^2 \times P^2} - \Phi \otimes \Phi \right)^{-1} \text{vec}(\Sigma), \quad (7.2.7)$$

□

Therefore, by the Property 7.2.1, given a coefficient matrix Φ and a covariance matrix Γ_0 , we can find the covariance matrix Σ of the error term Z_t such that the time series $\mathbf{X}$ generated by a VAR(1) process with the parameters Φ and Σ has a covariance matrix $\Gamma_X = \Gamma$. Such covariance matrix Σ is then given by:

$$\text{vec}(\Sigma) = \left(\text{Id}_{P^2 \times P^2} - \Phi \otimes \Phi \right) \text{vec}(\Gamma). \quad (7.2.8)$$

The coefficient matrices Φ for the three groups are chosen such that all the processes are stationary and relatively close to avoid a trivial classification:

$$\Phi_1 = \begin{pmatrix} 0.7 & 0.1 \\ 0.3 & 0.6 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0.65 & 0.1 \\ 0.2 & 0.6 \end{pmatrix}, \quad \Phi_3 = \begin{pmatrix} 0.6 & 0.1 \\ 0.3 & 0.6 \end{pmatrix} \quad (7.2.9)$$

The shared covariance matrix Γ is defined as:

$$\Gamma = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad (7.2.10)$$

For each group, we generate 100 time series of length $2^8 = 128$. The burn-in period of 1000 observations ensures that the initial conditions do not affect the properties of the generated data.

7.3 Frequency Domain

Time series need to be transformed in order to work in the frequency domain and utilize the methods developed in this thesis. Consequently, we firstly need to compute their estimated spectral densities.

For a VAR(1) process, the spectral density matrix $f(\omega)$ at frequency ω can be derived from the autoregressive parameters and the covariance matrix of the noise term. The spectral density matrix of a stationary time series $\mathbf{X}(t)$ following a VAR(1) process is given by the code from the thesis of J. Chau [8]. To visualize the spectral properties of the generated time series, we plot the true spectral density matrices for each class. This controlled case ensures that all classes have only one true spectral density, providing a clear benchmark for evaluating our classification methods.

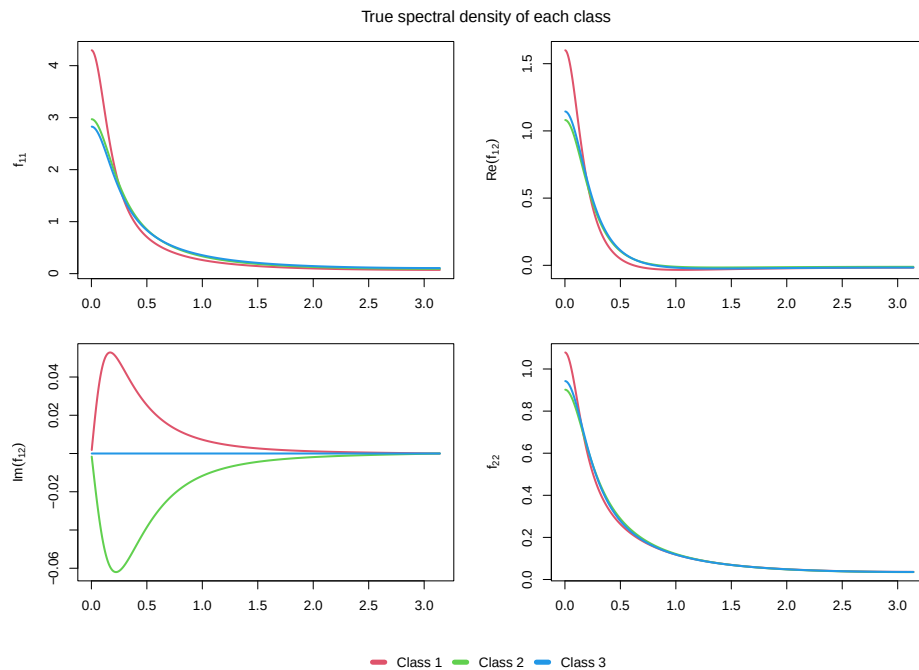


Figure 7.3.1: True spectral density matrices for the three classes of VAR(1) processes.

The true spectral densities depicted in Figure 7.3.1 demonstrate the distinct spectral characteristics of each class. This differentiation in the frequency domain underscores the importance of spectral density matrices for the classification task. More precisely, we can see that the auto-spectra f_{11} , f_{22} and the real part of the cross-spectra $\text{Re}(f_{12})$ of the different spectral density matrices of the classes are quite close, especially for classes 2 and 3. However, the imaginary part $\text{Im}(f_{12})$ is more differentiated according to the different classes. For the first class, $\text{Im}(f_{12})$ is positive; for the second class, it is negative; and for the third class, it is much closer to zero. Moreover, the imaginary part takes on smaller values than the real part, making it less robust in estimation. This suggests that differences in the

imaginary components may be less pronounced due to the lower power in the imaginary component of the cross-spectrum.

To estimate the spectral density matrices from the simulated time series, we employ the multitaper spectral estimator, defined in Definition 3.2.7, with the `pdPgram()` function from the `pdSpecEst` package [7]. The smoothing parameter B is crucial for ensuring the positive definiteness of the estimated spectral density matrices and needs to be at least greater than the time series dimension, here, $B \geq 2$. Moreover, we also have to select the time-bandwidth parameter n_w . Both parameters will be chosen by a cross-validation approach with all the classifiers, balancing smoothness and accuracy while maintaining sharpness at the most present frequencies. Finally, we select $B = 20$ and $n_w = 10$.

Although the theory suggests using periodograms for the estimated spectral densities of the observations when applying the Whittle distance, in practice, the results are significantly improved when these estimates are smoothed with the multitaper spectral estimator, especially when dealing with short time series. This approach not only enhances the reliability of the spectral estimates but also ensures better classification performance by reducing estimation variance and preserving the essential spectral characteristics of the data.

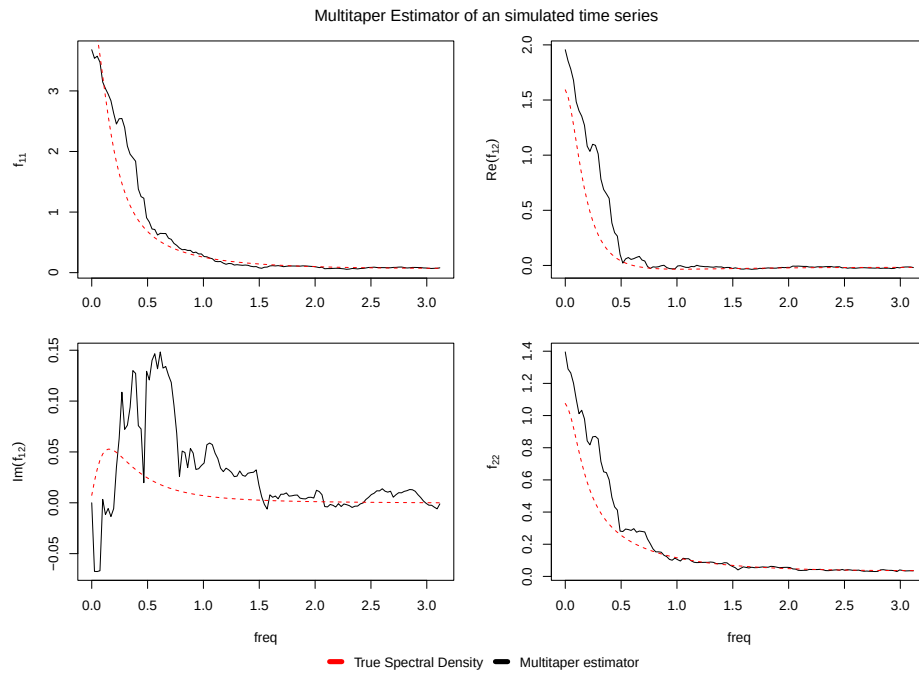


Figure 7.3.2: Multitaper spectral density estimate of an example VAR(1) time series. The plot shows the auto-spectra f_{11} and f_{22} , the real part of the cross-spectra $\text{Re}(f_{12})$, and the imaginary part $\text{Im}(f_{12})$ for the simulated time series.

Figure 7.3.2 shows an example of the estimated spectral density of a time series. The plot includes the auto-spectra f_{11} and f_{22} , the real part of the cross-spectra $\text{Re}(f_{12})$, and the imaginary part $\text{Im}(f_{12})$ for the simulated time series. It is evident that the imaginary

part of the cross-spectrum is less accurately estimated, likely due to the lower power in the imaginary component. This visualization helps in understanding the spectral properties and the differences between the classes, which are critical for effective classification.

By leveraging the frequency domain analysis, we aim to classify the generated time series accurately, utilizing the distinct spectral properties inherent in each class. This approach highlights the effectiveness of spectral density matrices in distinguishing between different types of time series data.

7.4 Classification

Now that the data are prepared and the estimated spectral densities are computed, we can proceed with the classification. We utilize two main classification methods: Minimum Distance to Means (MDM) and k -Nearest Neighbors (k -NN). These methods are combined with four different distance metrics: Euclidean, Log-Euclidean, Affine-Invariant Riemannian (AIR), and Whittle.

7.4.1 Classification Methods and Distance Metrics

K-Fold cross-validation To rigorously evaluate and obtain reliable results, we use K-Fold cross-validation. As shown in Figure 7.4.1, this method involves splitting the data into K subsets, or "folds." For each of the K iterations, one fold is used as the test set, while the remaining $K - 1$ folds are used as the training set. This process ensures that every observation is included in a test set exactly once and in a training set $K - 1$ times. Consequently, the test sets collectively form a partition of the original data, allowing the model to be evaluated on all data points. Furthermore, the training and test sets will maintain the same class distributions to avoid unbalanced datasets. For this application, we used 5 folds to evaluate the classifiers.

Now, let us briefly review the methods and the parameters needed for the different cases.

Minimum Distance to Means (MDM) : This method involves computing the mean spectral density matrix for each class. The mean is computed differently depending on the distance metric:

- **Euclidean and Whittle distances:** The classic Euclidean mean is used.
- **Log-Euclidean and AIR distances:** The mean is inherited from the geometry of the respective manifold, known as the Karcher mean.

For classification, the new observation is assigned the class of the closest mean spectral density matrix.

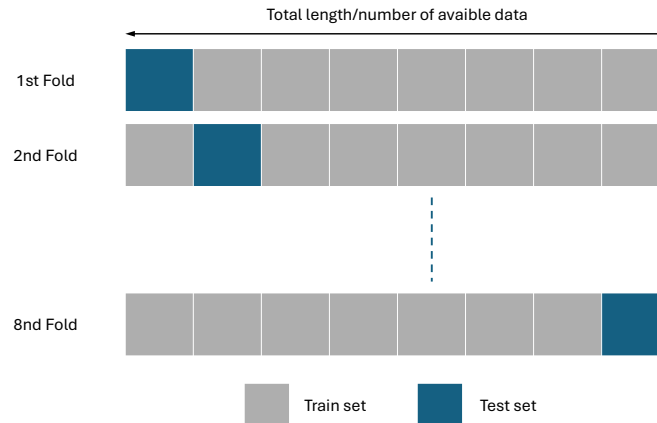


Figure 7.4.1: Example of Kfold cross-validation with 8 folds.

k-Nearest Neighbors (k -NN) : This method assigns a class to a new observation based on the majority class among its k -nearest neighbors in the spectral distance space. The same distance metrics used in MDM are applied here. The hyperparameter k , selected by cross-validation, determines the number of neighbors considered for classification, regulating the locality (local/global) of the comparison. We finally found $k = 10, 12, 18, 4$ respectively for the four distances: Euclidean, Log-Euclidean, Affine-Invariant Riemannian (AIR), and Whittle distance.

Finally, we have 8 possible classifiers using all the combinations of the 2 methods and the 4 distances. To evaluate them, we use accuracy as the metric, which is the proportion of observations that are correctly classified. The accuracy is the right metric in our case because we are dealing with balanced classes (100 observations in each class). However, if we worked with unbalanced classes, we would need to use other well-known metrics such as the micro F1-score or macro F1-score [27].

7.4.2 Results

The classification results are evaluated using the different methods and distance metrics. The optimal number of neighbors k for the k -NN method is determined through cross-validation.

Figure 7.4.2 shows the boxplot of the accuracies on the 5 folds obtained for the various combinations of classification methods and distance metrics. A dashed red line indicates the average results obtained by a dummy classifier, which classifies new observations randomly into the different classes.

The table 7.4.1 summarizes the mean accuracy obtained for all the combinations of the methods and the distances.

Upon examination of the results, several trends become apparent. The Euclidean distance consistently yields the lowest accuracy among all the distance metrics, particularly

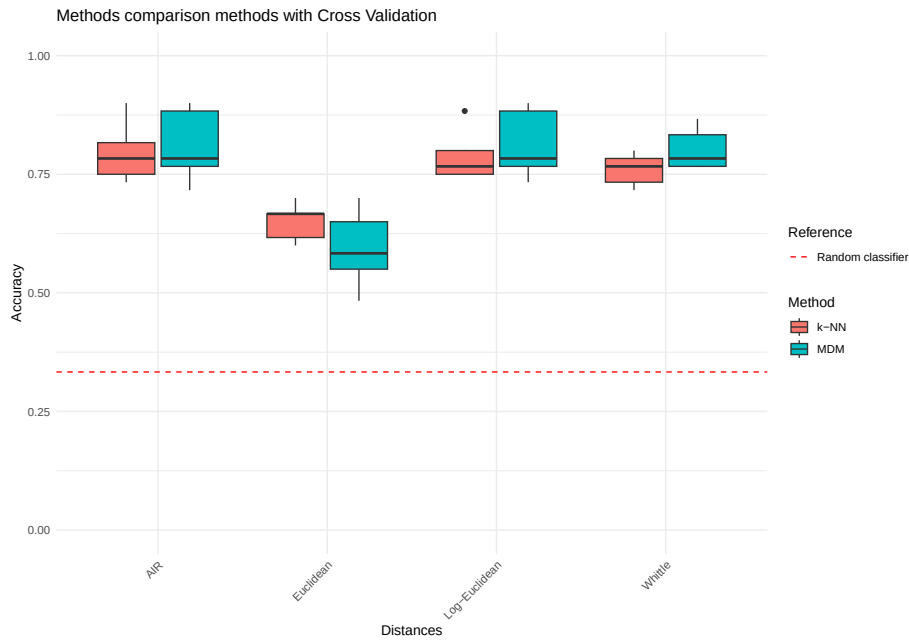


Figure 7.4.2: Boxplot of the accuracies on the 5 folds for different classification methods and distance metrics.

Method	AIR	Euclidean	Log-Euclidean	Whittle
k -NN	0.797	0.65	0.79	0.76
MDM	0.81	0.593	0.813	0.803

Table 7.4.1: Mean accuracy obtained for all the combinations of the methods and the distances.

when combined with the MDM method. Specifically, the average accuracy for the Euclidean distance is 0.65 when paired with k -NN and drops further to 0.593 when used with MDM. This suggests that the Euclidean distance, while providing a baseline for comparison, fails to capture the complex structure of the spectral density matrices as effectively as the other metrics.

In contrast, both the Log-Euclidean and Affine-Invariant Riemannian (AIR) distances deliver superior performances, with mean accuracies approaching 0.80. The Log-Euclidean distance achieves 0.79 with k -NN and 0.813 with MDM, while the AIR distance results in accuracies of 0.797 with k -NN and 0.81 with MDM. These results indicate that these two metrics are particularly well-suited for this classification task, offering robust and consistent performances across different classification strategies.

The Whittle distance also performs strongly, achieving an accuracy of 0.76 with k -NN and 0.803 with MDM. This shows that the Whittle distance is effective in this context, particularly when used with the MDM method. The strong performance of the Whittle distance with both k -NN and MDM methods suggests its relevance in capturing the relevant

spectral features of the time series data.

In conclusion, the classification performance is heavily influenced by the choice of distance metric. The Euclidean distance, while providing a straightforward baseline, is outperformed by the more sophisticated Log-Euclidean, AIR, and Whittle distances. Both the Log-Euclidean and AIR distances exhibit robust performance across the k -NN and MDM methods, making them reliable choices for spectral density matrix classification. The Whittle distance, while slightly less consistent than the other two, still delivers strong results, particularly with the MDM method, demonstrating its value in this domain.

Analysis of the Confusion Matrix To gain a deeper understanding of the performance of the best classifier —MDM with the AIR distance— we examine its confusion matrix, shown below. This matrix provides a detailed view of the classifier’s performance across the different classes.

		Actual		
		Class 1	Class 2	Class 3
Predicted	Class 1	96	2	0
	Class 2	3	62	26
	Class 3	1	36	74

Table 7.4.2: Confusion Matrix of the MDM with the AIR distance classifier.

The confusion matrix reveals that the classifier performs exceptionally well in distinguishing the first class, with 96 correct predictions out of 100, and only a small number of instances misclassified into the second class. However, the classifier struggles to distinguish between the second and third classes, as evidenced by the significant number of misclassifications between these two categories. Specifically, 36 instances that belong to the third class are misclassified as the second class, and 26 instances from the second class are incorrectly assigned to the third class.

This confusion between the second and third classes can be attributed to the inherent similarity in their spectral characteristics, as seen in the spectral density plots. Figure 7.3.1 shows that the spectral densities of these two classes are close to each other in the frequency domain, making it challenging for the classifier to differentiate between them.

Chapter 8

Application : Electroencephalography

In the previous section, we applied the described methods in this thesis to simulated data. This allowed us to create ideal conditions where all the underlying hypotheses of our methods were satisfied. For instance, we ensured that each class corresponded to a single true spectral density and that there were no outliers. However, real-world data rarely conforms to such perfect conditions. In this section, we will apply our methods to real data, acknowledging that the conditions are less controlled than in the simulated scenarios. The data chosen for this purpose are electroencephalography (EEG) data. The following repository contains the dataset used in this section: <https://github.com/sylvchev/dataset-ssvep-exoskeleton>, which was introduced in details in [23].

Electroencephalography (EEG) is a non-invasive method used to record electrical activity of the brain. EEG involves placing electrodes on the scalp to measure voltage fluctuations resulting from ionic current flows within neurons. These electrical signals are generated by the brain's millions of neurons that communicate with each other and with the body, controlling behavior and processing information. EEG captures these signals, providing insights into the brain's electrical activity, although it does not fully elucidate the complex workings of the brain.

Figure 8.0.1 shows a typical EEG setup with electrodes placed on the scalp according to the 10/20 system, a widely used method for electrodes placement.

EEG signals are characterized by different frequency bands that correspond to various brain activities. These frequencies are often associated with specific sensory and motor states such as eye movements, focus, and other cognitive functions. The ability to measure these frequencies opens up numerous applications, including brain-computer interfaces, medical diagnostics, and cognitive research.

Experimental Setup : For our study, we utilized Steady-State Visually Evoked Potentials (SSVEPs), a type of EEG response elicited by visual stimuli at specific frequencies. The EEG data were recorded using the g.Mobilab+ device at a sampling rate of 256 Hz across eight channels. The SSVEP stimulation was achieved using a flash stimulus technique with light-emitting diodes (LEDs) flashing at frequencies of 13, 17, and 21 Hz. To ensure precision, the LEDs were controlled by a microcontroller capable of flashing stimuli accurately to the millisecond.

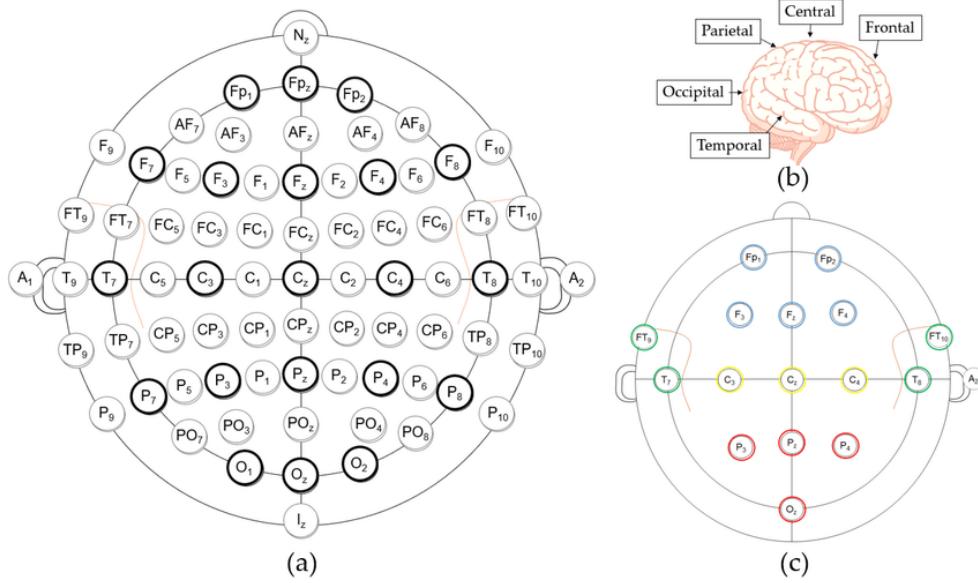


Figure 8.0.1: Typical EEG setup with electrodes placed on the scalp according to the 10/20 system.[17]

The electrodes were placed according to the 10/20 system at positions Oz, O1, O2, POz, PO3, PO4, PO7, and PO8, with the ground electrode placed on Fz and the reference electrode on the right or left mastoid.

Challenges and Considerations : Applying our methods to real EEG data introduces several challenges that were not present in the simulated data. One significant challenge is the variability between subjects. In cross-subject classification, the assumption that each class corresponds to a single true spectral density may be violated. Different subjects may exhibit different spectral characteristics in their EEG signals, even when responding to the same stimuli. This variability implies that the underlying processes are not identical across subjects, complicating the classification task.

8.1 Data Description

The dataset comprises records from 12 subjects, both male and female, aged between 20 and 28 years. Informed consent was obtained from all participants. Each subject was seated comfortably in an electric wheelchair during the recordings, with their right upper limb resting on an exoskeleton (which was not used during the experiments). A panel measuring 20x30 cm was attached to the left side of the chair, featuring three groups of four LEDs blinking at different frequencies. Despite the panel's position, the subjects could see it without moving their heads.

Participants were instructed to follow auditory cues indicating which LED to focus on or to fixate on a neutral point for the reject class. Each trial lasted five seconds, with a three-second pause between trials. Each session consisted of 32 trials, including eight trials

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Number of trials	64	64	64	64	64	64	96	64	64	128	64	160

Table 8.1.1: Total number of trials per subjects across all the sessions.

for each frequency (13 Hz, 17 Hz, and 21 Hz) and eight trials for the reject class. Records were saved in GDF format, with stimulation codes for each class available as time events. Each subject completed between two and five sessions on different days under consistent conditions. The total number of trials per subjects is shown in the Table 8.1.1

The authors of the dataset and the publications [23, 24, 22], noted that they excluded the following sessions for their study due to protocol issues (synchronization problem, hardware failure, etc): subject 08 session record-[2013.04.06-16.22.32], subject 10 sessions record-[2014.02.26-15.10.48] and record-[2014.02.26-15.50.09], subject 11 sessions record-[2014.02.24-17.56.37] and record-[2014.02.24-18.02.40].

Each trial initially lasted five seconds. However, the authors of [24] indicated that the best option is to remove the first two seconds after the stimulus. Thus, we have trials with a duration of three seconds at a 256 Hz sampling frequency, yielding a total length of $T = 768$ samples per trial. Each trial has eight channels, corresponding to the eight electrodes used for recording. We denote the k -th trial/observation of subject j as $\mathbf{X}_{(k,j)} \in \mathbb{R}^{768 \times 8}$. Furthermore, we denote the l -th class $\mathcal{C}_l$ respectively for the classes *Rest*, *13Hz*, *17Hz*, *21Hz*.

The structure of the dataset is summarized in the Figure 8.1.1 and is defined as follows:

- 12 subjects $S_1, \dots, S_{12}$.
- Each subject has M_{S_i} trials, where M_{S_i} is 64, 96, 128 or 160 trials. The Table 8.1.1 gives the exact number of trials per subject.
- Each subject's trials are equally distributed among the four possible classes $\mathcal{C}_k$ with $k = 1, \dots, 4$.
- Each class $\mathcal{C}_k$ of subject i has a representative denoted $\hat{\phi}_{k,i}$, which approximates the true spectral density $\phi_{k,i}$.

The Figure 8.1.2 shows an example of a trial. This trial was recorded on the first subject reacting to a 17Hz frequency signal. The neural activity captured by the 8 channels can be seen on the graph. Although the neural activity is different from one zone to another, we can observe some similarities between the channels. The spectral density will be the perfect tool to capture the relationship and dependency between the different channels and within each channel itself. It is anticipated that neuronal activity in the primary visual cortex will synchronize at the stimulus frequency when a subject fixes his gaze on a particular stimulus [24]. For instance, if a trial belongs to the class *17Hz*, which means a stimulus of 17Hz frequency is sent to the subject, then the neural activity should contain more power at the frequency of 17Hz. Thus, for each class, we expect to find a peak at the associated frequency in the spectral density matrix.

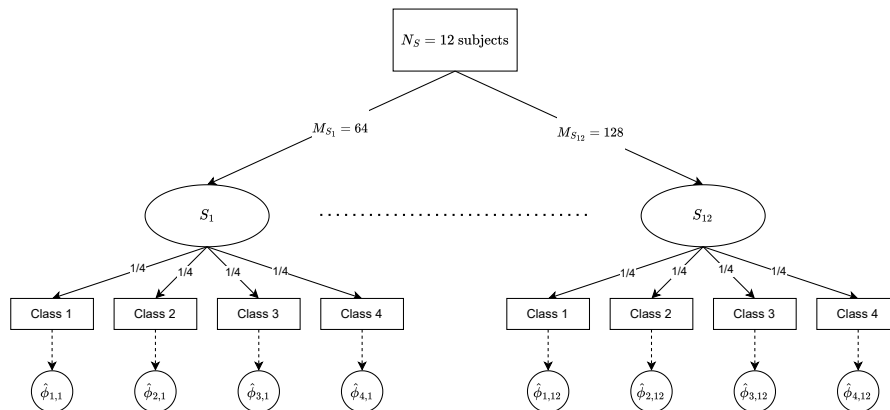


Figure 8.1.1: Structure of the dataset showing the distribution of trials across subjects and classes.

8.2 Frequency Domain Analysis

Similar to the simulation study, we start by computing the estimated spectral densities of the time series. To calculate the spectral density matrix of a time series, the time series must be weakly stationary. As illustrated in Figure 8.1.2, the data appears stationary, as the mean, variance, and autocovariances do not change over time. Similar observations can be made for other time series in the dataset. However, this visual assessment might not be sufficient to affirm the stationarity of the time series rigorously. It is advisable to conduct formal stationarity tests to validate this assumption.

The actual spectral density matrices, being unknown, are estimated using the multitaper spectral estimator, a smoothed periodogram, introduced in Definition 3.2.7. Both parameters and their values will be chosen through a cross-validation approach across all classifiers, aiming to balance smoothness and accuracy while preserving sharpness at the target frequencies of 13Hz, 17Hz, and 21Hz. This method ensures that the selected parameters optimize the performance of the spectral estimators in capturing the relevant frequency components. As we have discussed in the previous chapter, the same cross-validation technique was applied to determine the optimal parameters for the simulated time series, reinforcing the consistency of our methodology. After thorough testing, we select $B = 10$ and $n_w = 5$ as the optimal parameters, ensuring a reliable estimation of the spectral density matrices.

As previously outlined in the simulation study, although the theory suggests using periodograms for the estimated spectral densities of the observations when applying the Whittle distance, in practice, the results are improved when these estimates are smoothed with the multitaper spectral estimator.

Assumptions on True Spectral Densities In this analysis, we operate under the assumption that each class k of a subject i possesses a true spectral density represented as

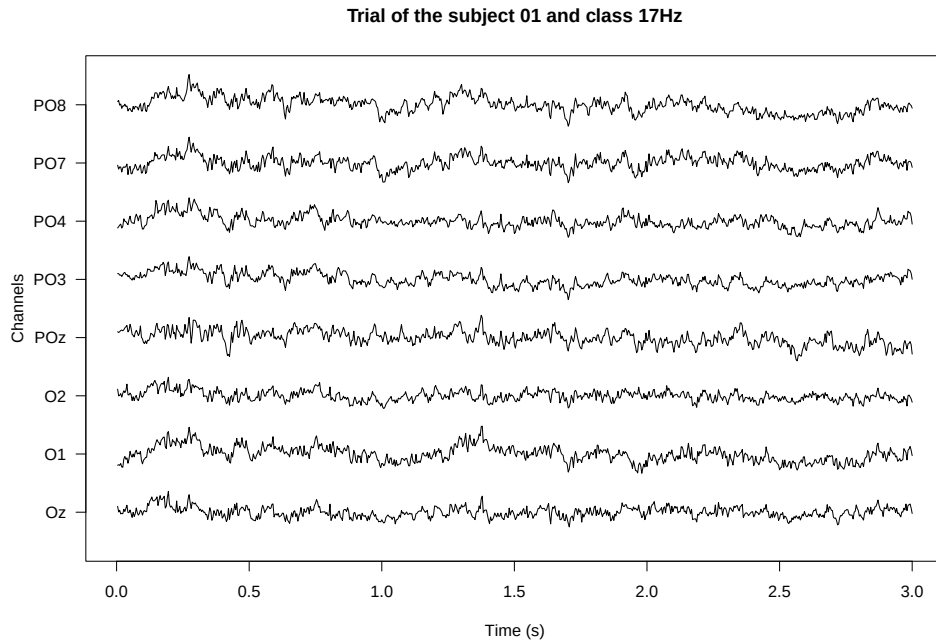


Figure 8.1.2: Example of a trial for the subject "01" and the class "17Hz".

$\phi_{k,i}$. This assumption is grounded in the idea that individual subjects exhibit consistent responses to identical stimuli, leading to stable spectral density characteristics. For instance, when exposed to the same auditory or visual stimulus, the neural response patterns within a specific subject tend to be uniform, resulting in a recognizable spectral density associated with that stimulus class.

However, it is crucial to note that two different subjects may exhibit divergent reactions to the same stimulus. This variability arises from inherent differences in individual neural processing, cognitive states, and physiological responses. As such, while subject i has a spectral density $\phi_{k,i}$ for class $\mathcal{C}_k$, subject j may have a different spectral density for the same class, denoted $\phi_{k,j}$. This leads to the general relationship:

$$\phi_{k,i} \neq \phi_{k,j}, \quad \text{for } i \neq j. \quad (8.2.1)$$

8.3 Classification

Now that the data are prepared and the estimated spectral densities are computed, we can proceed with classification.

- **Subject-Specific Classification :** This method involves training and testing the models separately on the observations of each individual subject. In other terms, there is one model per subject. The aim is to evaluate the model's performance specifically for each subject.

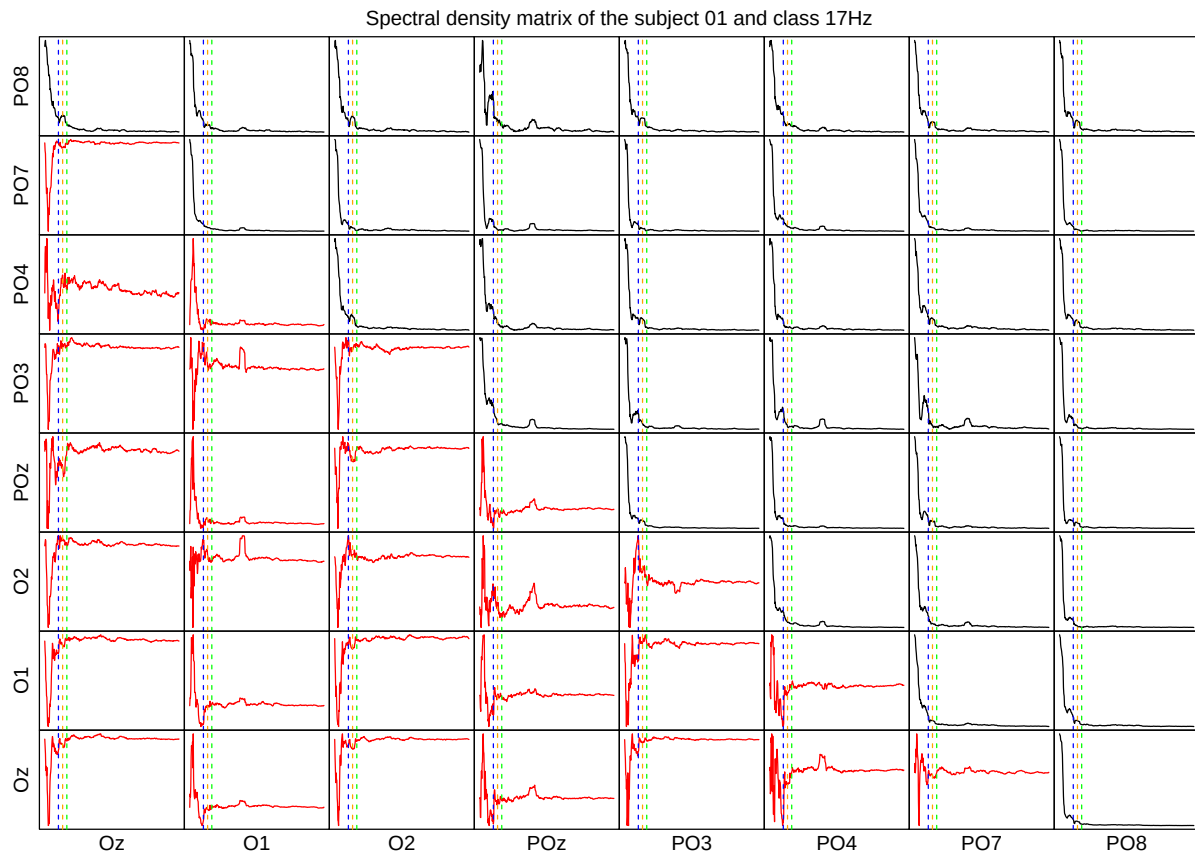


Figure 8.2.1: Spectral density matrix of the example trial in Figure 8.1.2. The imaginary part of the cross-spectra are in red, and the real part in black. The three dashed lines in blue, orange, and green represent the three frequencies associated with the different classes, i.e., 13Hz, 17Hz, and 21Hz.

- **Cross-subjects Classification:** In this approach, the models are trained on part of the subjects and then tested on the remaining subjects. In other terms, there is one model overall. This method provides insights into performance for a new subject in the dataset.

These two classification strategies offer different perspectives on model performance and can help assess the generalisation of the classification models in different contexts. The *Subject-Specific Classification* evaluation allows us to measure how good a model is at classifying new observations of an already known subject. The *Cross-Subjects Classification* evaluation allows us to measure how good a model is at classifying observations of a new subject.

To illustrate the difference between the two approaches, let us take the simple example illustrated in the Figure 8.3.1. Assume that we have observations from three subjects and that we split the data into train and test with a ratio of 2 : 1. In the Subject-Specific approach, we will train and test one model with the data of one subject. Therefore, each

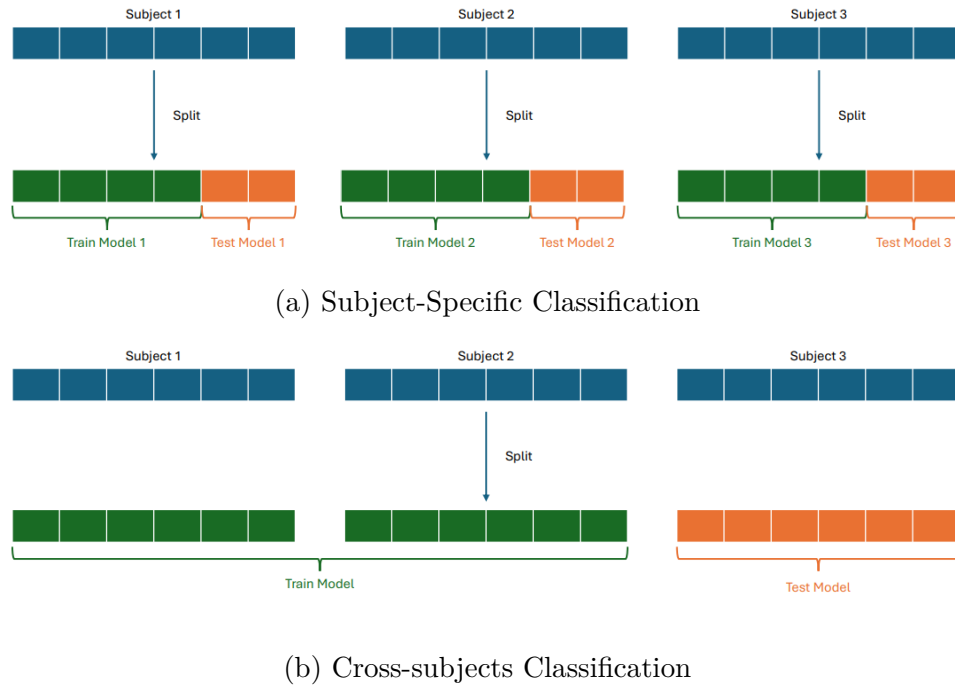


Figure 8.3.1: (a) Splitting approach and model for the Subject-Specific Classification, one model trained and tested per subject. (b) Splitting approach and model for the Cross-subjects Classification, one model trained and tested for all the subjects.

subject will be considered separately and we will have one model by subject. However, with the Cross-Subject approach, we will train one model with the data of 2 subjects, then test it on the data of the remaining subject.

For both classification approaches, we will proceed in the same way that described in the section 7.4.1 from the Simulation Study.

Frequency range used in distances : All the distances defined in the Chapter 6 allow us to compare two spectral densities on all frequencies, or on a subset of chosen frequencies, denoted Ω . Choosing all the frequencies can be a bad idea, because the signal can be noisy, such as EEG, and hide the information needed for the classification. Consequently, it is crucial to properly choose the range of frequencies that we have to consider. Generally, this can be tedious to find the optimal interval/subset due to the number of possibilities. This can be done by pre-selecting different intervals to try and by selecting the optimal with a cross-validation process. However, in this case, knowing the biological reaction expected with these stimuli, we know that the frequencies of interest are 13Hz, 17Hz and 21Hz. To capture the information around these frequencies, we choose to take a frequency range of $\Omega = [10\text{Hz}, 25\text{Hz}]$.

Generalization Capabilities of AIR distance : We also expect the AIR (Affine-Invariant Riemannian) distance to perform better due to its congruence invariance property 6.4.4. This property is particularly beneficial because it makes the Riemannian framework more robust to variations in the EEG source spatial distributions observed across different sessions and subjects. The congruence invariance, as demonstrated by [36], ensures that the distance between two observed spectral density matrices remains the same even if the mixing matrix changes.

In our context, the mixing matrix represents how the true neural signals (sources) are linearly combined and recorded by the EEG electrodes (observed signals). Different subjects or sessions might have different mixing matrices due to variations in electrode placement, head shape, or other physiological factors. Despite these differences, the AIR distance maintains consistent distance measures, thereby enhancing the classification performance across different subjects.

For instance, let us consider the spectral density matrices f_X and f_Y from two trials of a session and $A * f_X$ and $A * f_Y$ from another session, where A is a mixing matrix. The AIR distance ensures that $d_{AIR}(f_X, f_Y) = d_{AIR}(A * f_X, A * f_Y)$. This invariance helps maintain consistent distance measures despite variations in the sessions and subjects, described by the mixing matrices, leading to better generalization performance.

8.3.1 Subject-Specific Classification

The Subject-Specific Classification is expected to return a more accurate model than Cross-Subject Classification, because there is one classifier per subject. Indeed, as one model is trained on the data of only one subject assuming that a subject always reacts in the same way, we really have a true spectral density for each class. And this is one of the assumption for the MDM model.

To ensure a rigorous evaluation and obtain robust results, we use K-Fold cross-validation. As shown in Figure 7.4.1, this method involves splitting the data (in this case, data from a single subject) into K subsets, or "folds." For each of the K iterations, one fold is used as the test set, while the remaining $K - 1$ folds are used as the training set. This process ensures that every observation is included in a test set exactly once and in a training set $K - 1$ times. Consequently, the test sets collectively form a partition of the original data, allowing the model to be evaluated on all data points.

In the MDM method, as we proceed at a subject level, we need to estimate for each subject i the true spectral density for each class, denoted $\phi_{k,i}$ for the class $\mathcal{C}_k$. To do that, for a chosen distance d , we compute a representative by taking the Karcher mean, defined in 4.1.3, of the set containing the multitaper spectral estimators of the observed EEG of the subject i in the train set that belongs to the class $\mathcal{C}_k$. For the class k of the subject i , its representative is defined by

$$\hat{\phi}_{k,i}(\omega) := \arg \min_{S \in \mathcal{H}_P^+} \sum_{l \text{ s.t. } \mathbf{X}_{(l,i)} \in \mathcal{C}_k} d^2(S, \hat{f}_{(l,i)}(\omega)), \quad \forall \omega \in [-\pi, \pi], \quad (8.3.1)$$

where $\mathbf{X}_{(l,i)}$ is the l -th trial/observation of subject i and $\hat{f}_{(l,i)}$ its multitaper spectral

estimator. Then, we assign the observation of the test set to the same class than the closest representative class.

Even though Subject-Specific Classification operates on a per-subject basis, outliers can still occur due to intra-subject variability across different sessions. For example, if EEG data is collected from a subject over multiple sessions on different days, factors such as slight variations in electrode placement, changes in the subject's mental state, or external environmental influences can lead to inconsistencies in the recorded signals. These inconsistencies can act as outliers within the subject's own data.

However, as discussed in Section 6.7, the Karcher mean, especially when used with the AIR and log-Euclidean distances, is more robust to such outliers. These distances help mitigate the impact of outliers, resulting in a more accurate estimation of the true spectral density for each class. Thus, even when some trials are outliers due to factors like inconsistent electrode placement, the model's classification performance remains reliable.

8.3.1.1 Results

Now, let us apply all the methods to the EEG dataset. The Kfolds cross-validation is used to evaluate the methods with 4 folds. Therefore, for each methods we trained and tested $4 \text{ folds} \times 12 \text{ subjects} = 48$ accuracies.

For the k -NN method, we determined the optimal number of neighbors, k , by testing all possible values from 1 to 20 and selecting the value that yielded the best results. The optimal values of k identified for the different distances were 6, 8, 8, and 3 for the Euclidean, AIR, Log-Euclidean, and Whittle distances, respectively.

The results of all the combinations taking the two methods of classification and the four distances can be observed in the Figure 8.3.2. The boxplot graph of accuracies provides a comprehensive overview of the performance of different classification methods and distance metrics. Let us analyze and compare the results in details.

Method	AIR	Euclidean	Log-Euclidean	Whittle
k -NN	0.5930	0.4581	0.5465	0.4041
MDM	0.7250	0.4714	0.6903	0.5928

Table 8.3.1: Mean accuracy obtained for all the combinations of the methods and the distances with the Subject-Specific approach.

We can see in Figure 8.3.2 and Table 8.3.1 that, generally, the MDM methods yield better performance than the k -NN for each distance metric. Specifically, the MDM method shows a marked improvement over the k -NN method, except for the Euclidean distance, where the difference is negligible. The average difference in performance between the MDM and k -NN methods is approximately 0.12, indicating a substantial advantage for the MDM method in most scenarios.

The distance metric that provided the best results for both classification methods is the Affine-Invariant Riemannian (AIR) distance, achieving an accuracy of 0.593 for k -NN

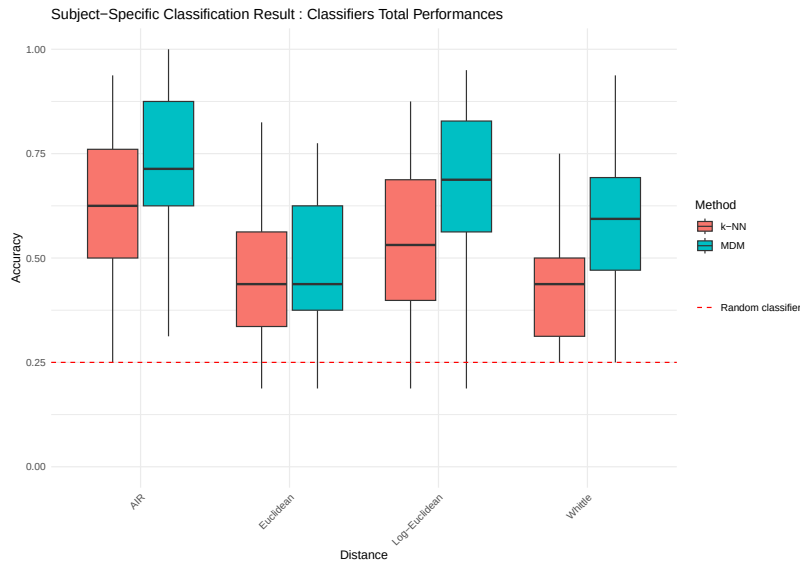


Figure 8.3.2: Boxplot of the total accuracies combining all the subjects results for each classifier. The graph compares the global performance of each classifier with the Subject-Specific approach.

and 0.725 for MDM. This highlights the robustness and effectiveness of the AIR distance in capturing the relevant features of the EEG data across different subjects. The superior performance of the AIR distance can be attributed to its congruence invariance property, which ensures consistent distance measurements despite variations in the data.

The best overall accuracy is given by the MDM method with the AIR distance, confirming its suitability for this type of data and classification task. The second most effective distance metric is the Log-Euclidean distance, followed closely by the Whittle distance.

In the case of the Whittle distance, as observed in the simulation study (Chapter 7), there is a significant difference between the MDM and k -NN performances. The MDM method achieves an accuracy more than 0.18 greater than the k -NN accuracy, confirming that the MDM method is better suited for leveraging the Whittle distance.

The Euclidean distance, however, shows the worst performance among all distance metrics. While it provides a baseline comparison, its simplicity does not capture the complex relationships in the data as effectively as the other distance metrics. The performance of the Euclidean distance is comparable to the Whittle distance when using the k -NN method, but both are significantly outperformed by the other combinations.

Despite the variations in performance, all classifiers outperform a random classifier, indicated by the red dashed line in Figure 8.3.2. This demonstrates that the methods and distance metrics used are effective in distinguishing between the different classes based on the EEG data.

In Figure 8.3.3, it is evident that the MDM method with AIR or Log-Euclidean distance consistently provides better classification performance across all subjects. The scatter plot shows that these methods achieve higher accuracies for nearly every subject, highlighting

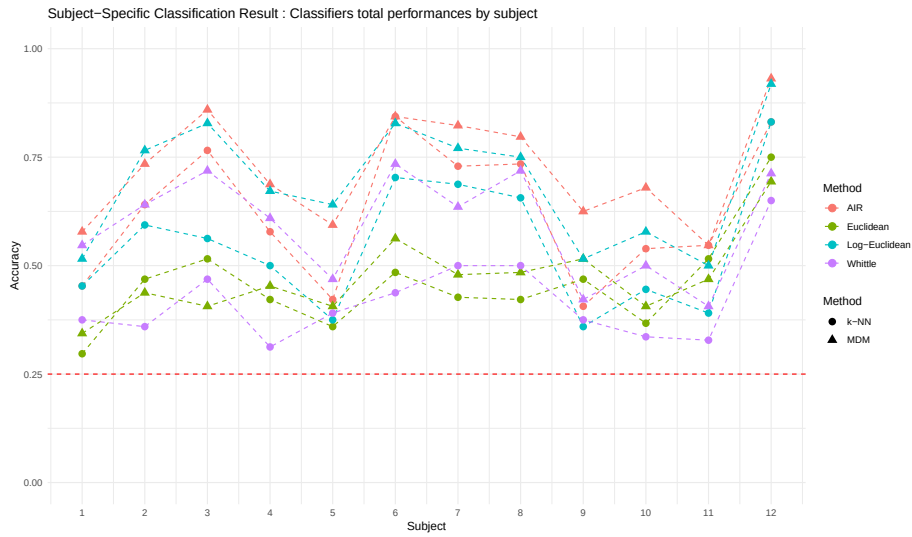


Figure 8.3.3: Accuracies by subjects obtained for each classifier. The graph compares the performance of each classifier per subject with the Subject-Specific approach. The red dashed line represents the average accuracy obtained by a random classifier.

their robustness and general applicability. Therefore, if we need to predict the stimulus based on the reactions of a new person, using the MDM method with AIR or Log-Euclidean distance will yield the best results. This consistency across subjects further validates the effectiveness of these methods and distance metrics in the subject-specific classification of EEG data.

8.3.2 Cross-Subject Classification

Cross-Subject Classification poses additional challenges compared to Subject-Specific Classification. One of the main issues is that the MDM model assumption, which relies on a single representative mean for each class, is violated. Indeed as explained in the paragraph 8.2 "Assumptions on True Spectral Densities" of the section 8.2, different subjects do not react in the same way to the same stimuli, resulting in variations within each class. Therefore, instead of having a true spectral density for each class, we can imagine that there are subclasses within each class, one for each subject.

For instance, we can see in the Figure 8.3.4 that the representative of the class 4 for subject 1 might differ significantly from that of subject 3. This variability complicates the classification process, as the model's representative of each class is approximated by averaging the responses from all subjects within that class.

In Cross-Subject Classification, we aim to evaluate the generalization power of the models to new subjects. The data are split intuitively: the training set contains part of the data from all subjects, and the test set contains the remaining part. This split is typically done using K-fold cross-validation, where the ratio of training to testing data are $(K - 1) : 1$. This method helps in assessing how well the model can generalize to new, unseen subjects.

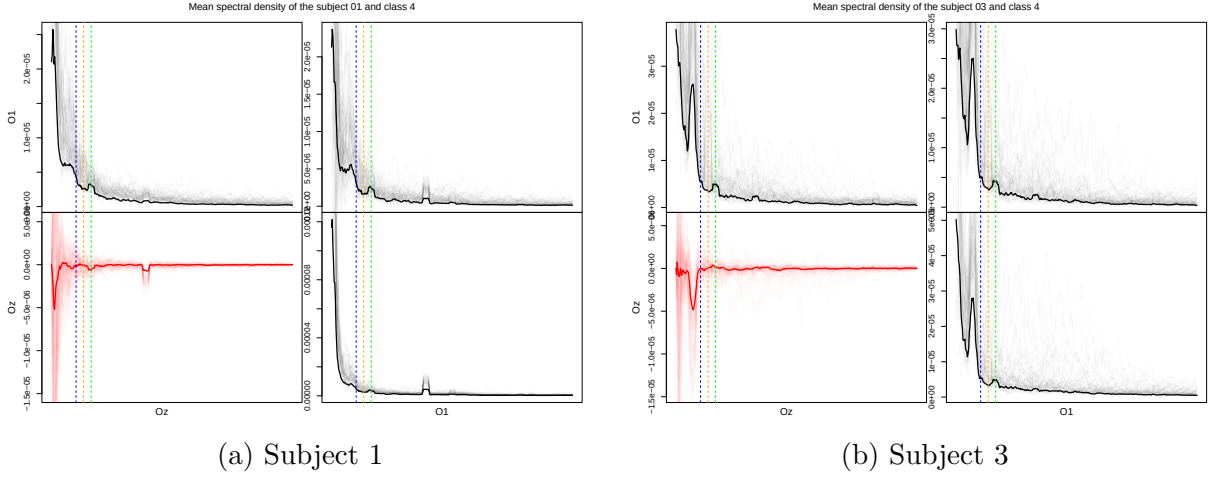


Figure 8.3.4: Comparison of the representatives of the fourth class 4 of the subject 1 and 3. Here, the representatives are the Karcher mean with the AIR distance and only the two first channels, Oz and O1, are displayed. The opaque line is the representative of the class while the transparent lines are the observations. The three dashed lines in blue, orange, and green represent the three frequencies associated with the different classes, i.e., 13Hz, 17Hz, and 21Hz.

In the MDM method, conversely to the Subject-Specific Classification, we do not proceed at a subject level, but with all subjects at the same time. Therefore, we need a common class representatives for all subjects, denoted ϕ_k for the class $\mathcal{C}_k$. To do that, for a chosen distance d , we compute a representative by taking the Karcher mean, defined in 4.1.3, of the set containing the multitaper spectral estimators of the observed EEG of all subjects in the train set that belong to the class $\mathcal{C}_k$. For the class k , its representative is defined by

$$\hat{\phi}_k(\omega) := \arg \min_{S \in \mathcal{H}_P^+} \sum_{l \text{ s.t. } \mathbf{X}_{(l,i)} \in \mathcal{C}_k} d^2(S, \hat{f}_{(l,i)}(\omega)), \quad \forall \omega \in [-\pi, \pi], \quad (8.3.2)$$

where $\mathbf{X}_{(l,i)}$ is the l -th trial/observation of subject i and $\hat{f}_{(l,i)}$ its multitaper spectral estimator. Then, we assign the observation of the test set to the same class as the closest representative class.

The heterogeneity in subjects' reactions poses significant challenges, particularly for the MDM method. For instance, if the majority of subjects exhibit response R_1 to stimuli while a minority exhibit response R_2 , the class representative derived during training will predominantly reflect R_1 . Consequently, if a test subject exhibits response R_2 , their observations may not align closely with the class representative, leading to potential misclassification.

This issue is less problematic for the k -NN method. If the test subject exhibits response R_2 , their observations are classified based on proximity to the nearest neighbors in the training set. Thus, the test subject's response is compared to other subjects who also exhibit response R_2 , resulting in more accurate classification.

Outliers can significantly impact the effectiveness of the MDM method in Cross-Subject Classification. If the training data used to construct the class representatives includes EEG records from a subject whose reactions deviate markedly from those of other subjects, these records may act as outliers. However, as demonstrated in Section 6.7, the Karcher mean associated with the AIR and log-Euclidean distances exhibits greater robustness to outliers. Consequently, employing these distances is expected to yield more reliable results in the presence of outliers, thereby enhancing the overall accuracy of the Cross-Subject Classification model.

8.3.2.1 Results

Now, let us apply all the methods to the EEG dataset. As we chose 12 folds for the cross-validation, i.e. tested only one subject in each fold, we have 12 accuracies for each methods.

With this classification, the optimal number of neighbors k are greater given that there are more observation to be compared in the training set. We determined the optimal k by testing all possible values from 1 to 50 and selecting the one that yielded the best results. The optimal values of k identified for the different distances were 7, 45, 32, and 15 for the Euclidean, AIR, Log-Euclidean, and Whittle distances, respectively.

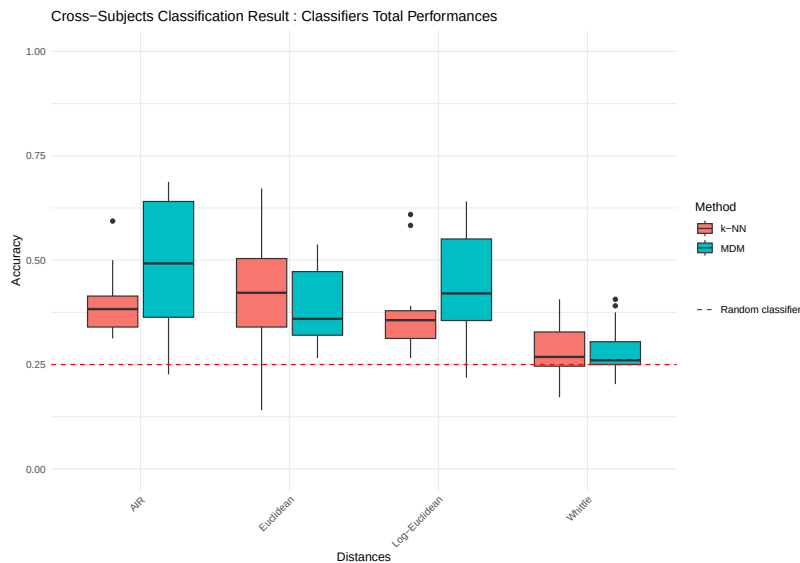


Figure 8.3.5: Boxplot of the total accuracies combining all the subjects results for each classifier. The graph compares the global performance of each classifier with the Cross-Subjects approach.

The results of the Cross-Subject Classification, presented in Figure 8.3.5 and Table 8.3.2, reveal several key insights. Generally, the performance of the MDM method is better than the k -NN method for most distances, similar to the trend observed in the Subject-Specific

Method	AIR	Euclidean	Log-Euclidean	Whittle
k -NN	0.3971	0.4189	0.3788	0.2778
MDM	0.4842	0.3816	0.4377	0.2856

Table 8.3.2: Mean accuracy obtained for all the combinations of the methods and the distances with the Cross-Subjects.

Classification. However, the overall accuracies are lower due to the increased variability between subjects.

The AIR distance, which performed the best in the Subject-Specific Classification, continues to show strong results in the Cross-Subject Classification, achieving an accuracy of 0.4842 for MDM and 0.3971 for k -NN. The AIR distance's congruence invariance property contributes to its robustness, making it effective even when there are variations in the subjects' reactions to the same stimuli.

The Euclidean distance, surprisingly, shows a higher accuracy for k -NN (0.4189) compared to MDM (0.3816). This could be due to the simplicity of the Euclidean distance, which might be more forgiving in scenarios with high inter-subject variability.

The Log-Euclidean distance performs relatively well, with accuracies of 0.4377 for MDM and 0.3788 for k -NN. This distance metric captures the geometric structure of the spectral density matrices, providing a balance between complexity and performance.

The Whittle distance shows the lowest performance, with accuracies of 0.2856 for MDM and 0.2778 for k -NN. Consequently, knowing that a random classifier has a average accuracy of 0.25, we can conclude that this distance is not adapted for this task.

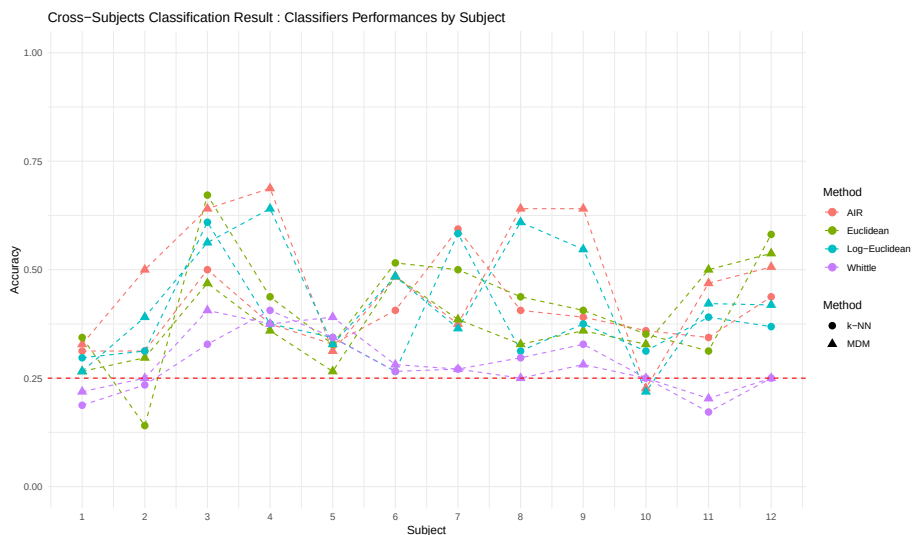


Figure 8.3.6: Accuracies by subjects obtained for each classifier. The graph compares the performance of each classifier per subject with the Cross-Subjects approach. The red dashed line represents the average accuracy obtained by a random classifier.

In Figure 8.3.6, we can see the accuracies for each subject across different classifiers. Unlike in the results of the Subject-Specific Classification, the scatter plot that no single method consistently outperforms the others across all subjects. Indeed, we can observe that, depending on the subject, one method often outperforms the other, regardless of the distance used. For instance, for subjects 2, 4, 8, 9, and 11, the MDM method yields the best classifiers. Conversely, for subjects 7 and 10, the k -NN method provides better classifiers. This discrepancy can be attributed to the fact that subjects 7 and 10 exhibit unique reactions to stimuli, which deviate significantly from the average reaction. Consequently, their responses are quite different from the class representatives calculated in the MDM method. On the other hand, subjects 2, 4, 8, 9, and 11 have more typical reactions that are closer to the mean response, making them better represented by the MDM method.

Overall, these results indicate that while Cross-Subject Classification is inherently more challenging than Subject-Specific Classification, certain methods and distance metrics, such as the MDM with AIR distance, still manage to achieve relatively good performance. This reinforces the importance of selecting appropriate distance metrics and classification methods tailored to the specific challenges posed by cross-subject variability.

8.4 Comparison with Previous Research

In 2020, Clément Vaes' master thesis at UCLouvain [46] explored a similar dataset but employed a different classification approach. Vaes utilized a more parametric method, where he filtered the time series at specific target frequencies—13Hz, 17Hz, and 21Hz. This filtering process resulted in time series with dimensions of 3 frequencies $\times$ 8 channels = 24 dimensions. He then calculated the covariance matrix of these filtered time series, contrasting it with our approach of comparing the estimated spectral densities.

Vaes also conducted dimensionality reduction, which improved the accuracy of his classifiers. He applied the same methods as in our study, specifically k -Nearest Neighbors (k -NN) and Minimum Distance to Mean (MDM).

For the Subject-Specific Classification, both Vaes and our study achieved the best results using the AIR distance and the MDM method. Vaes reported an average accuracy of 0.7083 without dimensionality reduction, whereas our approach yielded a slightly higher average accuracy of 0.725. Moreover, after applying dimensionality reduction, Vaes achieved an even higher average accuracy of 0.75.

Vaes' findings underscore the potential benefits of dimensionality reduction for EEG data classification. However, it is important to note that his method relies on the prior knowledge of target frequencies, making it less generalizable compared to our more comprehensive approach.

8.5 Conclusion & Discussion

In the Subject-Specific Classification of EEG data, the tailored approach of training a model for each individual subject yields superior performance compared to the Cross-Subject Classification. This subject-specific approach capitalizes on the consistency of each subject's neural responses, enabling more accurate estimations of the spectral density for different classes. The K-fold cross-validation strategy employed ensures robust evaluation and the mitigation of overfitting, thereby producing reliable accuracy metrics.

Our findings underscore the effectiveness of the Minimum Distance to Mean (MDM) method, particularly when paired with advanced distance metrics such as the Affine-Invariant Riemannian (AIR) and Log-Euclidean distances. These metrics demonstrate a significant ability to manage outliers and intra-subject variability, which are common challenges in EEG data analysis. The MDM method consistently outperforms the k -Nearest Neighbors (k -NN) classifier across different distance metrics, with the AIR distance yielding the highest classification accuracies.

In contrast, the Euclidean distance exhibits the poorest performance, underscoring its inadequacy in handling the complex, non-Euclidean structure of the EEG data. This is particularly evident when comparing the performance of the MDM method across different metrics, where the Euclidean distance lags significantly behind AIR and Log-Euclidean distances.

The consistency of the MDM method's superior performance across subjects suggests its potential as a reliable tool for subject-specific EEG classification tasks. The results indicate that for practical applications, particularly those requiring high accuracy and robustness against variability and outliers, the MDM method in combination with AIR or Log-Euclidean distances is recommended.

When extending to Cross-Subject Classification, however, the challenge of inter-subject variability becomes prominent. The heterogeneity in neural responses across different subjects complicates the classification process, reducing the effectiveness of models that rely on common class representatives. No method is universally optimal, and the effectiveness of a given approach may vary depending on the specific circumstances. This necessitates a careful consideration of the strengths and weaknesses of each method in relation to the particular context of use. Despite the robustness of the Karcher mean, the variance in performance across subjects remains high, suggesting that further refinement is necessary to improve generalization in Cross-Subject Classification scenarios.

A promising direction for future research is the incorporation of dimensionality reduction techniques, such as selective channel analysis, which could help to refine the feature space and improve classification accuracy. For instance, by identifying and focusing on the most informative EEG channels, particularly those less susceptible to noise or variability due to electrode placement differences, the classification models could become more robust. This approach aligns with the observations from prior studies, where dimensionality reduction has shown to enhance classifier performance in EEG analysis.

Overall, while Subject-Specific Classification using the MDM method with advanced distance metrics offers a reliable and robust approach, Cross-Subject Classification poses

additional challenges that require further investigations. Future work should explore strategies to enhance the generalization capability of models in the presence of inter-subject variability, potentially through adaptive or hybrid classification techniques, as well as through the strategic reduction of dimensionality in the feature space.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

This thesis aimed to contribute to the advancement of the field of time series classification by examining spectral densities through the lens of Riemannian geometry. The primary objective was to investigate and evaluate advanced distance metrics for comparing spectral densities, with an additional focus on assessing the performance of various classification methods.

The research demonstrated that spectral densities, when analyzed within a Riemannian geometric framework, provide a robust and effective approach to improving time series classification. In particular, the usage of advanced distance metrics, including the Affine-Invariant Riemannian (AIR), Log-Euclidean, and Whittle distances, proved to be markedly more efficacious than the conventional Euclidean distance. In particular, the Affine-Invariant Riemannian (AIR) and Log-Euclidean distances demonstrated superior performance, especially in real-world datasets, where their resilience to outliers offers a considerable advantage. Empirical studies have demonstrated remarkable robustness in the presence of outliers, thereby enhancing the practical applicability of these methods.

Furthermore, the comparative analysis of classifiers, including Minimum Distance to Means (MDM) and k -Nearest Neighbors (k -NN), revealed the considerable impact that different spectral distances have on classification accuracy. This finding underscores the importance of selecting distance metrics that are well-suited to the specific characteristics of the data, thereby optimizing the classification process.

Nevertheless, the research was limited to stationary time series, indicating that future work should extend these methods to address non-stationary time series, thereby broadening their applicability across diverse datasets.

9.2 Discussion and Future Work

The findings of this thesis open several avenues for further research in time series classification. While the advancements achieved provide a strong foundation, there are numerous

opportunities to enhance and expand upon this work.

Optimizing Dimensionality Reduction: One significant area for future research involves optimizing the selection of dimensions within classifiers. As discussed in Chapter 8, the work by Vaes [46] suggests that refining the selection of dimensions could further improve classifier accuracy. This approach could lead to more precise and reliable classifications by better weighting and utilizing different features.

Exploring Weighted Spectral Densities: Another promising direction is the application of weighted spectral densities. Li et al. (2013) [29] have indicated that assigning different weights to dimensions in distance calculations can enhance classifier performance. Developing a weighted distance metric tailored to the specific characteristics of time series data could be highly beneficial. This could involve applying a Hermitian and positive definite (HPD) weighting matrix $\mathbf{W}$ to the spectral densities, preserving the HPD manifold's geometric structure and allowing the continued use of the tools developed in this thesis. The challenge will be to find an optimal method for selecting the weighting matrix $\mathbf{W}$.

Theoretical Validation of AIR Distance Superiority: Further investigation is necessary to formally ascertain the theoretical superiority of the AIR distance over the Euclidean distance. Although the empirical results presented in this thesis indicate that the AIR distance enhances time series classification, a rigorous theoretical analysis could provide a deeper understanding of the underlying reasons for this phenomenon. By investigating the mathematical characteristics of the AIR distance, particularly its capacity to preserve the geometric structures of spectral densities, we could substantiate its status as a preferred metric and justify its wider implementation in diverse applications. Such a foundation would not only reinforce the conclusions of this research but also make a substantial contribution to the field of time series analysis.

Investigating Non-Stationary Time Series: In many real-world applications, the assumption of stationarity in time series analysis can be overly restrictive. Time series data, such as EEG recordings during a brain seizure, often exhibit spectral characteristics that change over time, reflecting dynamic underlying processes. Traditional stationary models may fail to capture these evolving patterns, leading to inaccuracies in classification. By relaxing the stationarity assumption, we can explore more flexible models that allow for time-dependent spectral characteristics. This approach, as suggested by the work of Joris Chau in his 2018 thesis [8], emphasizes the need for local time-frequency analysis to better understand and classify time series with non-stationary behaviors. By adopting nonparametric methods for locally stationary time series, we can improve the accuracy and robustness of time series classification in real-world, dynamic environments.

Robustness of the Karcher Mean: Finally, the robustness of the Karcher mean associated with AIR and Log-Euclidean distances warrants further theoretical examination.

The empirical results of this thesis suggest that these distances offer significant benefits, particularly in the presence of outliers. Investigating the robustness of the Karcher mean, using methods such as calculating the influence function or breakdown point, see [19], could provide deeper insights into the stability and reliability of these Karcher means.

Optimizing MDM for Multimodal Distributions: Another optimization of the MDM method can be explored in the context of multimodal distributions, particularly in Cross-Subject Classification. As discussed in Section 8.3.2, different subjects may exhibit varying reactions to the same stimulus, leading to a scenario where a single class, such as Class A, could actually contain multiple subclasses representing distinct types of reactions. Throughout this thesis, we have often treated the minority of subjects with less frequent reactions as outliers. However, these "outliers" could instead be considered as forming one or more distinct subclasses within Class A. When these heterogeneous reactions are grouped together, the representative of Class A could be skewed, resulting in a poor representation, as illustrated in Section 5.3. To address this issue, a clustering method could be applied prior to classification to identify these subclasses. By working with more homogeneous subclasses, the resulting representatives would be more accurate and reliable, thereby improving the overall performance of the MDM method.

In conclusion, while this thesis has made substantial contributions to time series classification, further research in these areas could lead to more sophisticated and practical solutions. Continuous refinement and expansion of these methods could further advance time series analysis, ultimately resulting in more accurate and insightful predictions across a variety of applications.

Bibliography

- [1] Hadi Ratham Al Ghayab, Yan Li, Siuly Siuly, and Shahab Abdulla. Epileptic EEG signal classification using optimum allocation based power spectral density estimation. *IET Signal Processing*, 12(6):738–747, 2018. __eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-spr.2017.0140>.
- [2] Gerold Alsmeyer. Chebyshev’s Inequality. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 239–240. Springer, Berlin, Heidelberg, 2011.
- [3] T. Ando, Chi-Kwong Li, and Roy Mathias. Geometric means. *Linear Algebra and its Applications*, 385:305–334, July 2004.
- [4] David R. Brillinger. *Time Series: Data Analysis and Theory*. 1981.
- [5] Jorge Caiado, Nuno Crato, and Daniel Peña. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10):2668–2684, June 2006.
- [6] Jorge Caiado, Nuno Crato, and Pilar Poncela. A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification*, 14(1):117–146, March 2020.
- [7] Joris Chau. pdSpecEst: An Analysis Toolbox for Hermitian Positive Definite Matrices, January 2020.
- [8] Van Vinh Chau. *Advances in spectral analysis for multivariate, nonstationary and replicated time series*. PhD thesis, UCL - Université Catholique de Louvain, 2018.
- [9] D. S. Coates and P. J. Diggle. Tests For Comparing Two Estimated Spectral Densities. *Journal of Time Series Analysis*, 7(1):7–20, 1986. Publisher: Wiley Blackwell.
- [10] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. Publisher: Taylor & Francis.
- [11] James W. Cooley and John W. Tukey. An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 19(90):297–301, 1965. Publisher: American Mathematical Society.

- [12] Peter J. Diggle and Nicholas I. Fisher. Nonparametric Comparison of Cumulative Periodograms. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(3):423–434, 1991. Publisher: [Royal Statistical Society, Oxford University Press].
- [13] P. Thomas Fletcher and Sarang Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, February 2007.
- [14] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. Advanced Spectral Methods for Climatic Time Series. *Reviews of Geophysics*, 40(1):3–1–3–41, 2002. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2000RG000092>.
- [15] Brian Hall. The Matrix Exponential. In Brian C. Hall, editor, *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, pages 31–48. Springer International Publishing, Cham, 2015.
- [16] Scott H. Holan, Christopher K. Wikle, Laura E. Sullivan-Beckers, and Reginald B. Cocroft. Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals. *Biometrics*, 66(3):914–924, September 2010.
- [17] Seunghyeok Hong and Hyun Baek. Drowsiness Detection Based on Intelligent Systems with Nonlinear Features for Optimal Placement of Encephalogram Electrodes on the Cerebral Area. *Sensors*, 21:1255, February 2021.
- [18] Hsiao-Yun Huang, Hernando Ombao, and David S. Stoffer. Discrimination and Classification of Nonstationary Time Series Using the SLEX Model. *Journal of the American Statistical Association*, 99(467):763–774, 2004. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
- [19] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, September 2011. Google-Books-ID: j1OhquR_j88C.
- [20] Alexis Ioannou, Konstantinos Fokianos, and Vasilis J. Promponas. Spectral density ratio based clustering methods for the binary segmentation of protein sequences: A comparative study. *Biosystems*, 100(2):132–143, May 2010.
- [21] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019.
- [22] Emmanuel Kalunga, Sylvain Chevallier, and Quentin Barthélemy. Data augmentation in Riemannian space for Brain-Computer Interfaces. In *STAMMLINS*, STAMMLINS - ICML Workshop on Statistics, Machine Learning and Neuroscience, Lille, France, June 2015.

- [23] Emmanuel Kalunga, Sylvain Chevallier, Olivier Rabreau, and Eric Monacelli. Hybrid interface: Integrating BCI in multimodal human-machine interfaces. In *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pages 530 – 535, Besançon, France, July 2014.
- [24] Emmanuel K. Kalunga, Sylvain Chevallier, Quentin Barthélemy, Karim Djouani, Eric Monacelli, and Yskandar Hamam. Online SSVEP-based BCI using Riemannian geometry. *Neurocomputing*, 191:55–68, May 2016.
- [25] Edmund Landau. *Handbuch der Lehre von der Verteilung der Primzahlen*. Leipzig B.G. Teubner, 1909.
- [26] John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2018.
- [27] Kenneth Leung. Micro, Macro & Weighted Averages of F1 Score, Clearly Explained, September 2022.
- [28] Yili Li and Kon Wong. Signal classification by power spectral density: An approach via Riemannian geometry. pages 900–903, August 2012.
- [29] Yili Li and Kon Max Wong. Riemannian Distances for Signal Classification by Power Spectral Density. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):655–669, August 2013. Conference Name: IEEE Journal of Selected Topics in Signal Processing.
- [30] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, May 2015.
- [31] Jason Lines, Anthony Bagnall, Patrick Caiger-Smith, and Simon Anderson. Classification of Household Devices by Electricity Usage Profiles. In Hujun Yin, Wenjia Wang, and Victor Rayward-Smith, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, pages 403–412, Berlin, Heidelberg, 2011. Springer.
- [32] Hugo Daniel Macedo and José N. Oliveira. Typing linear algebra: A biproduct-oriented approach, December 2013. arXiv:1312.4818 [cs, math].
- [33] Derek Hylton Maling. *Coordinate systems and map projections*. Pergamon Press, Oxford[etc], 2nd ed edition, 1992. OCLC: 911279490.
- [34] Paul A. Parker, Scott H. Holan, and Nalini Ravishanker. Nonlinear Time Series Classification Using Bispectrum-based Deep Convolutional Neural Networks, March 2020. arXiv:2003.02353 [cs, stat].
- [35] Ofer Pasternak, Nir Sochen, and Peter J. Basser. The effect of metric selection on the analysis of diffusion tensor MRI data. *NeuroImage*, 49(3):2190–2204, February 2010.

-
- [36] Xavier Pennec and Nicholas Ayache. Uniform Distribution, Distance and Expectation Problems for Geometric Features Processing. *Journal of Mathematical Imaging and Vision*, 9(1):49–67, July 1998.
 - [37] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66, January 2006.
 - [38] Peter Petersen. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer New York, 2006.
 - [39] Rainer von Sachs. Nonparametric Spectral Analysis of Multivariate Time Series. *Annual Review of Statistics and Its Application*, 7(Volume 7, 2020):361–386, March 2020. Publisher: Annual Reviews.
 - [40] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer International Publishing, Cham, 2017.
 - [41] Christopher A. Sims. Macroeconomics and Reality. *Econometrica*, 48(1):1–48, 1980. Publisher: [Wiley, Econometric Society].
 - [42] D. Slepian. Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—V: The Discrete Case. *Bell System Technical Journal*, 57(5):1371–1430, 1978. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1978.tb02104.x>.
 - [43] Dayang Sun. Predictive Modeling of Seismic Events: A Comparative Analysis Of K-Nearest Neighbors and Random Forest Algorithms. *Highlights in Science, Engineering and Technology*, 85:272–279, March 2024.
 - [44] Adam M Sykulski, Sofia C Olhede, Arthur P Guillaumin, Jonathan M Lilly, and Jeffrey J Early. The debiased Whittle likelihood. *Biometrika*, 106(2):251–266, June 2019.
 - [45] David Thomson. Spectrum Estimation and Harmonic Analysis. *Proceedings of the IEEE*, 70:1055–1096, October 1982.
 - [46] Clément Vaes. Classification of biological signals on the symmetric positive definite manifold. Master’s thesis, UCLouvain - Université Catholique de Louvain, EPL, 2020.
 - [47] Lian van der Krieke, Frank J. Blaauw, Ando C. Emerencia, Hendrika M. Schenk, Joris P. J. Slaets, Elisabeth H. Bos, Peter de Jonge, and Bertus F. Jeronimus. Temporal Dynamics of Health and Well-Being: A Crowdsourcing Approach to Momentary Assessments and Automated Generation of Personalized Feedback. *Psychosomatic Medicine*, 79(2):213, March 2017.

-
- [48] P. Whittle. Estimation and information in stationary time series. *Arkiv för Matematik*, 2(5):423–434, August 1953. Publisher: Institut Mittag-Leffler.
 - [49] Xinru Yuan, Wen Huang, Pierre-Antoine Absil, and Kyle A. Gallivan. Averaging Symmetric Positive-Definite Matrices. In Philipp Grohs, Martin Holler, and Andreas Weinmann, editors, *Handbook of Variational Methods for Nonlinear Geometric Data*, pages 555–575. Springer International Publishing, Cham, 2020.

