CIMAT

Centro de Investigación en Matemáticas, A.C.

Detection of changes in time series: a frequency domain approach

Tesis

Que para obtener el Grado de:

Doctorado en Ciencias con Orientación en Probabilidad y Estadística

P R E S E N T A:

Carolina de Jesús Euán Campos

Director:

Dr. Joaquín Ortega Sánchez

Guanajuato, Guanajuato, México

16 Agosto de 2016

Integrantes del Jurado.

Presidente: Dra. Graciela María de los Dolores González Farías
CIMAT
Secretario: Dr. Rolando José Biscay Lirio
CIMAT
Vocal: Dr. Hernando Ombao
Universidad de California en Irvine
Vocal: Dr. Gabriel Rodríguez Yam
Universidad Autónoma de Chapingo
Vocal y director de tesis: Dr. Joaquín Ortega Sánchez
CIMAT
Lector especial: Dr. Pedro César Alvarez Esteban
Universidad de Valladolid

Asesor:

Dr. Joaquín Ortega Sánchez.

Sustentante:

M.C. Carolina de Jesús Euán Campos.

To my parents, Bolivar and Maria Concepción.

To Israel M Hdz.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr Joaquín Ortega, for his patience, suggestions and guidance during the last four years.

I am grateful to my external examiner, Prof Hernando Ombao, for his advices and suggestions to this project. I wish to thank Prof Pedro C. Álvarez Esteban for several fruitful conversations during the last years and his comments on this thesis.

A special note of thanks is also extended to the committee members of oral defense, Dra Graciela González, Dr Rolando Biscay and Dr Gabriel Rodríguez, for their time dedicated to comment and to make suggestions to the thesis.

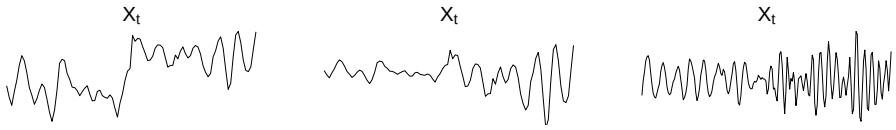
I thank CIMAT for the facilities to carry out this research and CONACYT for the PhD scholarship given.

Contents

Chapter 1

Introduction

In time series analysis, the stationarity assumption is fundamental. However, for real applications this assumption is not always satisfied, due to changes in the process over time. A change point τ is a time point when the probability distribution of a time series changes. A process $\{X_t\}$ can change in different ways, for example a change in mean or a change in variance (see for instance panel (a) and (b) in Figure 1.1). These cases have been studied by many researchers and different statistical tests have been developed, some of which are based on sample moments and computing algorithms.



(a) Change in mean (b) Change in variance (c) Change in spectra

Figure 1.1: Processes that change in (a) mean, (b) variance and (c) spectra.

Our interest lies in considering changes in spectra (see panel (c) in Figure 1.1). A change in spectra means a change in the waveforms of the signal and it is very important in many applications. For example, during a storm sea waves become higher and slower, this has many implications in the construction of maritime structures. Another example is the study of brain signals taken by electroencephalograms, in this case an activation of a brain

region shows faster oscillations of the signal (i.e. more transfered energy in shorter periods of time). In both cases, the understanding of where and how these changes happen is relevant. A large part of the literature on spectral analysis is based on the stationarity assumption for the process. However, in some cases, we need to carry out spectral analysis for processes that are not stationary. There are several points of view from which an analysis of this sort can be assessed. The most frequently used is the detection of change points in the process. The main assumption of this approach is that a process $\{X_t\}$ changes in k specific time points, $\tau_1, \tau_2, \ldots, \tau_k$, where both the number and location of the change points are unknown and the process is assumed to be stationary between them. Another approach is to model the process $\{X_t\}$ as a locally stationary process.

This project considers, as an alternative to the change point approach, the use of clustering methods for time series to identify periods or segments that have similar spectra. Time series clustering has captured the attention of many researchers in the past few years.

1.1 Previous Results

1.1.1 Spectral theory for a stationary process

Let us consider the following definition of stationarity.

Definition 1.1. *A time series $\{X(t),\ t \in \mathbb{R}\}$, is said to be stationary if for all t, s and r in $\mathbb{R}$ (i) $\mathbb{E}|X(t)|^2 < \infty$, (ii) $\mathbb{E}(X(t)) = m$, and (iii) $cov(X(t+s), X(t+r))$ does not depend on t.*

We will denote by $\gamma(h) = Cov(X(t), X(t+h))$, the covariance function of the stationary time series $\{X(t)\}$.
The basis of time series spectral analysis are Herglotz's theorem and the Spectral Representation theorem. The proofs can be found in different sources such as Brockwell and Davis (2006) and Shumway and Stoffer (2011).

Theorem 1.1 (Herglotz's Theorem)**.** A complex-valued function $\gamma(\cdot)$ defined on the integers is non-negative definite, i.e., $\sum_{i,j=1}^{n} a_i \gamma(i-j) a_j \geq 0$ for all positive integers n and all vector $\mathbf{a} \in \mathbb{R}^n$, if and only if

$$\gamma(h) = \int_{-1/2}^{1/2} e^{i2\pi\omega h} \mathrm{d}F(\omega), \tag{1.1}$$

for all $h = 0, 1, \ldots$, where $F(\cdot)$ is a right-continuous, non-decreasing, bounded function on $[-1/2, 1/2]$ and $F(-1/2) = 0$.

F is called the *spectral distribution* function of γ, and if $F(\omega) = \int_{-1/2}^{\omega} f(\nu)\mathrm{d}\nu$, then f is called the *spectral density* of γ. In terms of the time series $X(t)$, if $\gamma(\cdot)$ is absolutely summable, the spectral density is the Fourier transform of the covariance function, i.e.,

$$f(\omega) = \sum_{-\infty}^{\infty} \gamma(h) e^{-i2\pi\omega h}, \qquad -\frac{1}{2} \leq \omega \leq \frac{1}{2}.$$

Theorem 1.2 (The Spectral Representation Theorem)**.** If $\{X_t\}$ is a stationary sequence with mean zero and spectral distribution F, then there exists a right-continuous orthogonal-increment process $\{Z(\omega)\}$ such that

(i) $\mathbb{E}|Z(\omega) - Z(-1/2)|^2 = F(\omega)$,

(ii) $X_t = \int_{-1/2}^{1/2} e^{i2\pi\omega t} \mathrm{d}Z(\omega)$.

The spectral representation can be interpreted as follows, "any stationary time series can be looked at as a sum of infinitely many cosine and sine waveforms with random coefficients" [Shumway and Stoffer (2011)]. At lag $h = 0$ one gets $\gamma(0) = \mathrm{Var}(X_t) = \int_{-1/2}^{1/2} f(\omega)\mathrm{d}\omega$. Thus, the time series variance is decomposed over the frequency domain, where the spectrum at frequency ω can be roughly interpreted as the variance contributed by the oscillation in a narrow frequency band around ω.

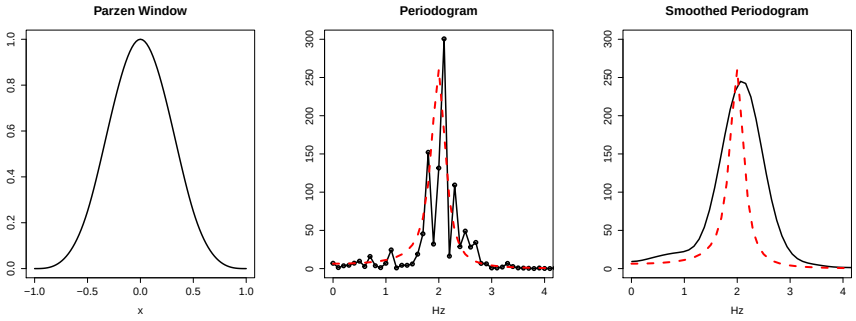
The spectral analysis of a time series can be seen from a nonparametric or parametric approach. For example, in the case of the parametric ARMA(p,q) model, the spectral density has a closed form as

$$f(\omega) = \sigma^2 \frac{|\theta(e^{i2\pi\omega})|^2}{|\phi(e^{i2\pi\omega})|^2},$$

where θ and ϕ are the p^{th} and q^{th} degree polynomials, $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$.

Based on the stationary assumption, a natural nonparametric estimator for the spectral density is the periodogram which is defined as

$$I(\omega_j) = n^{-1} \left| \sum_{t=1}^{n} x_t e^{-it2\pi\omega_j} \right|^2 = \sum_{|h|<n} \hat{\gamma}(h) e^{-it2\pi\omega_j},$$



(a) (b) (c)

Figure 1.2: Estimation of the spectral density, the true density is the dashed red curve. (a) Parzen window used in the lag window estimator. (b) Estimator using the periodogram. (c) Estimator using the smoothed lag window

for a stationarity and centered time series X_t, $t = 1, \ldots, T$, at the fundamental Fourier frequencies $\omega_k = k/T$, $k = 1, \ldots, n$ with $n = \lfloor (T-1)/2 \rfloor$, where $\hat{\gamma}$ is the sample autocovariance function. This estimator is asymptotically unbiased but the variance does not go to 0 as T increases. For this reason, it is common to smooth the periodogram, we will work with the smoothed lag window estimator,

$$\hat{f}(\omega) = \sum_{|h|<a} \beta(h/a)\hat{\gamma}(h)e^{-it2\pi\omega_j}, \tag{1.2}$$

where $\beta(x)$ is an even, piecewise continuous function of x satisfying the conditions: 1) $\beta(0) = 1$, 2) $|\beta(x)| \leq 1$ for all x, and 3) $\beta(x) = 0$ for $|x| > 1$. Figure 1.2 shows an example of these estimators. The spectral density is symmetric, i.e, $f(\omega) = f(-\omega)$; hence, we use the one sided spectra, which is defined as $f^*(\omega) = 2f(\omega)$ with $0 < \omega \leq 1/2$.

1.1.2 Change point detection

An approach to study the changes in a process is to consider that the changes occur at specific time points and between two change points the process is stationary, i.e., considering the process as piecewise stationary.

Let $\{X_t\}$ be a process that changes in K time points, $\tau_1, \tau_2, \ldots, \tau_K$, where the number of change points and their locations are unknown. Let $f_j(\omega)$ be

the spectrum of the kth segment, i.e., the spectrum of $\{X_t\}$ for $\tau_k \leq t < \tau_{k+1}$. Davis et al. (2006) developed a methodology called Auto-PARM where the main idea is to fit an AR model to each piece. The minimum description length principle is applied to find the "best" combination of the number of segments, the lengths of the segments, and the orders of the piecewise AR processes. The estimates are strongly consistent, however, approximating a nonstationary time series by a parametric AR model which may not be reasonable in some applications.

Another parametric framework is the Detection of Changes by Penalized Contrasts (DCPC) developed by Lavielle (1999) and collaborators. They considered a sequence of real random variables $\{X_t\}_{t=1,\ldots,n}$ and assumed that the distribution of the process depends on a parameter θ that changes abruptly at some unknown instants $\{\tau_i, 1 \leq i \leq K\}$, where K is also unknown. To estimate both K and the change points $\{\tau_i, 1 \leq i \leq K\}$ they used a penalized contrast function of the form $J(\mathbf{t}, \mathbf{y}) + \beta pen(\mathbf{t})$, where the contrast function is defined as

$$J(\mathbf{t}, \mathbf{y}) = \sum_{k=1}^{K} C(X_{\tau_{k-1}}, \ldots, X_{\tau_k}),$$

and $pen(\mathbf{t})$ is a penalization term and β is a tuning parameter. The contrast is the addition of local contrast functions at each segment $[\tau_{k-1}, \tau_k]$. The contrast function at each segment depends on the estimator $\hat{\theta}(X_{t_{k-1}}, \ldots, X_{t_k})$ calculated on the kth segment of $\mathbf{t}$ which is defined as the solution of a minimization problem of a function U. A particular case of this methodology has been studied in Lavielle and Ludeña (2000). They considered a parametric spectral density $f(\omega; \theta)$, and chose the function U as the Whittle log likelihood between the periodogram and the parametric model. So, $\hat{\theta}_k$ is the Whittle estimator in the kth segment. Finally, the contrast function J is taken as a weighted average of the Whittle log likelihood of each segment. Simulation studies show that the DCPC method performs well when K is known. In the case of K unknown, how to determine the penalization term is not clear. This election will be taken subjectively in most of the cases.

The methodologies mentioned before explore the change point detection as a minimization problem Another approach to study the change point problem is as a statistical hypothesis test, as follows. We would like to test, for a given segment j, the hypothesis

$$H_0 : \ f_j(\omega) = f_{j+1}(\omega) \ \forall \omega \qquad vs \qquad H_A : \exists \omega_0 \text{ such that } f_j(\omega_0) \neq f_{j+1}(\omega_0).$$

Dette and Paparoditis (2009) and Dette and Hildebrandt (2012) studied this hypothesis test. The test statistic proposed is a functional of the euclidean norm of the difference between the spectra integrated over $[-\pi, \pi)$. To establish a rejection region they proposed two options; 1) it can be proved that the test statistic is asymptotically normally distributed or 2) a bootstrap procedure based on a Wishart distribution. The power of the test is good in the examples shown in the paper. The asymptotic distribution is dependent on the smoothing of the periodogram.

Two other possible statistics are studied in Jentsch and Pauly (2012) with special interest in the case of unequal length time series ($n_1 < n_2$). The statistics are based on the periodogram,

$$T_n^{(1)} = \frac{1}{n_2} \sum_{j=1}^{n_2} c_j (I_{n_1,X}(\omega_j) - I_{n_2,Y}(\omega_j))^2$$

and

$$T_n^{(2)} = \frac{1}{n_2} \sum_{j=1}^{n_2} c_j \log^2 \left(\frac{I_{n_1,X}(\omega_j)}{I_{n_2,Y}(\omega_j)} \right) \mathbb{1}_{\{I_{n_1,X}(\omega_j) I_{n_2,Y}(\omega_j) \neq 0\}}.$$

In both cases an asymptotic distribution is obtained. In the first case, it converges to a random variable which is a linear combination of a sequence of independent double exponentially distributed random variables. The second statistic converges to a random variable which is a linear combination of a sequence of independent standard logistic distributed random variables.

The asymptotic distribution of $T_n^{(1)}$ depends on unknown values of the spectra which need to be estimated. Hence, an asymptotically exact test can not be applied. The asymptotic distribution of $T_n^{(2)}$ has the advantage of being distribution-free under H_0, but the power of the test is low. As a promising way to improve the performance of these statistics, the authors proposed, as future work, to use integrated periodograms or smoothed periodograms as estimators of the spectrum.

1.1.3 Spectral theory for a locally stationary process

The spectral analysis for locally stationary process was developed by Dahlhaus (1997).

Definition 1.2. *A sequence $X_{t,T}$ $(t = 1, \dots, T)$ is called locally stationary with transfer function A^0 and trend μ if there exists a representation*

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t) A^0_{t,T}(\lambda)\, d\xi(\lambda), \tag{1.3}$$

where the following holds.

(i) $\xi(\lambda)$ is a stochastic process on $[-\pi, \pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ and

$$cum\{d\xi(\lambda_1), \dots, d\xi(\lambda_k)\} = \eta\left(\sum_{j=i}^{k} \lambda_j\right) g_k(\lambda_1, \dots, \lambda_k) d\lambda_1 \cdots d\lambda_k,$$

where $cum\{\cdots\}$ denotes the cumulant of kth order, $g_1 = 0$, $g_2 = 1$, $|g_k(\lambda_1, \dots, \lambda_k)| \leq M_k$ (constant) for all k and $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + 2\pi j)$ is the 2π periodic extension of the Dirac delta function.

(ii) There exists a constant K and a 2π-periodic function $A : [0,1] \times \mathbb{R} \to \mathbb{C}$ with $A(u, -\lambda) = \overline{A(u, \lambda)}$ and

$$sup_{t,\lambda} \left| A^0_{t,T} - A\left(\frac{t}{T}, \lambda\right) \right| \leq K T^{-1}$$

for all T: $A(u, \lambda)$ and $\mu(u)$ are assumed to be continuous in u.

From the definition and other results (see Dahlhaus, 2011) it can be shown that for $u_0 \in [0, 1]$ there exists a stationary process $\tilde{X}_t(u)$ such that

$$|X_{t,T} - \tilde{X}_t(u)| = O_p\left(\left|\frac{t}{T} - u_0\right| + \frac{1}{T}\right),$$

which justifies the name "locally stationary process". $X_{t,T}$ has a unique time varying spectral density which is, locally, the same as the spectral density of $\tilde{X}_t(u)$. Furthermore, it has, locally, the same auto-covariance since $cov(X_{[uT],T}, X_{[uT]+k,T}) = c(u, k) + O(T^{-1})$ uniformly in u and k, where $c(u, k)$ is the covariance function of $\tilde{X}_t(u)$. This justifies taking $c(u, k)$ as the local covariance function of $X_{t,T}$ at time $u = t/T$. This suggests to estimate a local spectra with rolling windows.

A more formal estimation method for the time varying spectral density was proposed by Dahlhaus (2000), using an approximation to the Gaussian

likelihood. The proposed quasi likelihood is a generalization of the Whittle likelihood for a stationary process. The generalization is obtained using the localized periodogram or preperiodogram which uses only the pairs $X_{[t+0.5-k/2]}, X_{[t+0.5+k/2]}$ to estimate the covariance of lag k at time t. He looked at the parametric time-varying models and proved asymptotic properties for the resulting estimator.

Another methodology to estimate the time varying spectra was developed by Ombao et al. (2005). In this case, they work under the SLEX framework (smooth localized complex exponentials, which is a collection of orthogonal bases). They built a family of multivariate models that characterized the time varying spectra. To select the best model in this family a penalized log energy is used. The resulting method is flexible, computationally efficient, and easy to interpret.

The study of locally stationary processes has been a topic of research for many years. One of the challenges in this area is building a hypothesis tests for this family of processes. Sergides and Paparoditis (2008) established a general framework for hypothesis testing in the multivariate case. The main idea is to compare the spectra between the processes to answer a specific scientific question, for example, are the spectra equal? are the time series uncorrelated? etc. They assume a parametric model and use the L^2 norm for the test statistic. A bootstrap method is used to establish the rejection criterion. The examples of hypotheses are simple, a more complicated hypothesis could be more difficult to study using this framework.

Another important scientific question is whether a parametric model is a good option. Preuss et al. (2013) investigated the problem of testing semiparametric hypotheses in locally stationary processes. The best parametric estimator under the null hypothesis is chosen minimizing a local version of the Kullback-Leibler divergence,

$$\mathcal{L}(u,\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log g(\theta,\lambda) + \frac{I_N(u,\lambda)}{g(\theta,\lambda)}\right)\mathrm{d}\lambda,$$

between the parametric family and the periodogram. To establish the rejection criteria an asymptotic distribution of the test statistic and a bootstrap method is used. In general, the bootstrap method shows better results than using the asymptotic distribution.

Change point detection and the locally stationary approaches are not completely separated. Last and Shumway (2008) studied the problem of

detecting abrupt changes in a piecewise locally stationary time series. The goal was segmenting a nonstationary time series into locally stationary segments. The proposed method consists of comparing, with the symmetric Kullback-Leibler divergence, the right and left spectra at time t. Given a time t, compute the estimated spectral density $\hat{f}_L(t,\lambda)$ with $x_{t-n+1},\ldots,x_t$ and $\hat{f}_R(t,\lambda)$ with $x_{t+1},\ldots,x_{t+n}$, then obtain

$$D_1(t) = \frac{1}{n}\sum_{\lambda}\left(\frac{\hat{f}_L(t,\lambda)}{\hat{f}_R(t,\lambda)} - \frac{\hat{f}_R(t,\lambda)}{\hat{f}_L(t,\lambda)}\right).$$

Next, find the maximun of $D_1(t)$ and take

$$D_2(t) = \begin{cases} D_1(t) & if \quad x \in A^c \\ \\ 0 & if \quad x \in A \end{cases}$$

where $A = [t_i - n, t_i - 1] \cup [t_i + 1, t_i + n]$. Repeat until the values of D are sufficiently "small". To decide this critical value the asymptotic distribution or bootstrap methods are used.

Depending on the specific problem one prefers one approach over the other. For example, in the study of ocean waves the changes are slow, so methods to detect abrupt changes usually give poor results. Also, there could be transition periods between stationary states, which could be modelled within the locally stationary framework.

1.1.4 Time Series Clustering

In general, clustering is a procedure whereby a set of unlabeled data is divided into groups so that members of the same group are similar, while members of distinct groups differ as much as possible. The problem of clustering when the data points are time series has received a lot of attention in recent times. Liao (2005) gives a revision of the field up to 2005. A most recent review on time series clustering can be found in Caiado et al. (2015).

There exists a big variety of applications in different fields. Lachiche et al. (2005) developed a method for fMRI where the area between the variations of the signals around their means was used as similarity measure together with a Growing Neural Gas (GNG) clustering algorithm. Other examples are: the identification of similar physicochemical properties of amino acid sequences

(Savvides et al., 2008), detection of groups of stocks sharing synchronous time evolutions with a view towards portfolio optimization (Basalto and De Carlo, 2006), the identification of geographically homogeneous regions based on similarities in the temporal dynamics of weather patterns (Bengtsson and Cavanaugh, 2008) and finding groups of similar river flow time series for regional classification (Corduas, 2011), to name a few.

According to Liao (2005) there are three approaches to time series clustering: methods based on the comparison of raw data, feature-based methods, where the similarity between time series is gauged through features extracted from the data and methods based on parameters from models adjusted to the data. We are interested in the second group using the spectral density of the corresponding time series as the principal feature, i.e., the time series will be characterized by its spectral density and produce groups with signals having similar spectral densities. Then, the problem of detection of changes has been transformed to an equivalent problem, finding similarities in the behavior of time series.

We will use the clustering method to identify periods with similar spectral density and hence consider them as a bigger stationary period. However, the proposal will be more general and could be also used to identify similarities in space instead of time.

The rest of the thesis is organized as follows. In **Chapter 2** the concept of the total variation (TV) distance is introduced. The TV distance is used to quantify the similarity between spectral densities. Previous approaches have in common the use of a similarity measure between spectra. The most frequently used are the L^2 norm and the Kullback-Leibler divergence, however, it wil be shown that these may not be adecuate measures to detect small changes. In **Chapter 2** the properties of the TV distance between estimated spectra and its capacity to detect small changes are established.

In **Chapter 3** the proposed methodologies which belong to the time series clustering approaches are given. Two different proposals, the TV distance in a hierarchical clustering method and the Hierarchical Spectral Merger (HSM) method are considered. In both cases the feature of interest for clustering is the spectral density and the TV distance is used as a measure of similarity. However, there are important differences in the clustering procedures that make both methods distinct. Simulation studies show that the first proposal is more convenient to detect slow changes. The HSM

method has the advantage of having an accurate procedure to choose the number of clusters, based on the distribution of the TV distance.

Finally, in **Chapter 4** we present applications to two different cases of study. The first is related to the study of ocean waves, where the main goal is the identification of stationary periods. We use the first proposal in this case. The second belongs to the neuroscience framework, in particular the analysis of electroencephalogram (EEG) data. The main goal is the detection of activation or anomalies in any channel during the resting state of a motor skill experiment. The HSM method is used in this example.

Chapter 2

Total Variation Distance

To detect changes in spectra, we need a quantity that gauges gauges the extent of similarity between two spectral densities. Our proposal is to use the total variation (TV) distance as a similarity measure to compare spectral densities. The TV distance is defined in general for any two probability measures, we shall adapt it to the case of spectral densities. A statistical application of the TV distance was proposed by Alvarez-Esteban et al. (2012), where a hypothesis test to compare two probability densities is developed.

Definition 2.1. *Let $(\mathcal{X}, \mathcal{A})$ be any measurable space. The total variation distance d_{TV} between two probability measures P and Q on $\mathcal{X}$ is defined as*

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

In our case, we will use the total variation distance on the real line, then $\mathcal{X} = \mathbb{R}$ and $\mathcal{A} = \mathbb{B}(\mathbb{R})$, where $\mathbb{B}(\mathbb{R})$ is the class of the Borel sets on the real line.

An important property of the TV distance is that it is bounded between 0 and 1. This property can be easily deduced from the definition. A value of 1 for the distance can be attained if P and Q have disjoint support. This property is very useful in order to interpret distances: values close to 1 mean that the two measures are quite different while distance values close to 0 mean that they are very similar, almost equal. In terms of spectral densities, if the TVD is equal to 1 then the spectral content of the two signals are completely different, i.e., they not share a common frequency band.
If P and Q have density functions (typically with respect to the Lebesgue measure, μ), f and g, the TV distance between them can be computed using

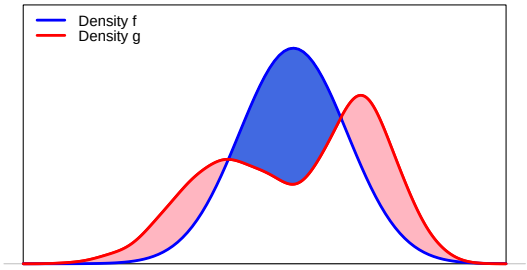


Figure 2.1: The TV distance measures the similarity between the two densities. The blue shaded area, that is equal to the pink, is the value of the TV distance.

the following expression,

$$d_{TV}(P, Q) = 1 - \int \min(f, g)\ d\mu.$$

This equation helps to graphically interpret the TV distance. If two densities f and g, have TV distance equal to $1 - \alpha$ this means that they share a common area of size α. Figure 2.1 illustrates the case with two density functions, the area of the pink or blue regions represents the TV distance. Both colored regions represent the non-common part of the density functions, while the white area under the curves is the common part.

2.1 TV distance and the Wasserstein distance

The total variation distance has an important interpretation in the framework of contamination models (Alvarez-Esteban et al., 2012), that could be extended to the spectral analysis of time series. To interpret this notion of contamination in the case of spectral densities, we shall define the general concept of a contamination model.

A contamination model for a probability P_0 consists of assuming that P_0 cannot be observed because the presence of noise, or contamination of level $\varepsilon \in (0, 1)$, which follows a probability law N, so that we observe

$$P = (1 - \varepsilon)P_0 + \varepsilon N.$$

If one only considers two measures, the similarity between them is defined as follows.

Definition 2.2. *Two probability measures P and Q on the same sample space are α-similar if there exist probability measures λ, P', and Q' such that*

$$\begin{aligned} P &= (1-\varepsilon_1)\lambda + \varepsilon_1 P' \\ Q &= (1-\varepsilon_2)\lambda + \varepsilon_2 Q' \end{aligned} \tag{2.1}$$

with $0 \leq \varepsilon_i \leq \alpha,\ i = 1, 2$.

Smaller values of α correspond to more similar probability measures. Before we connect this concept with the total variation distance, we consider the case of measures on the real line and the definition of the Wasserstein distance.

Definition 2.3. *Given $\alpha \in (0, 1)$, we define the set of α-trimmed versions of P by*

$$\mathcal{R}_\alpha(P) := \left\{ Q \in \mathcal{P} : Q \ll P,\ \frac{dQ}{dP} \leq \frac{1}{1-\alpha} \right\},$$

where $\mathcal{P}$ denotes the set of Borel probability measures on $\mathbb{R}$. Or equivalently,

$$\mathcal{R}_\alpha(P) := \left\{ Q \in \mathcal{P} : Q \ll P,\ Q(A) \leq \frac{1}{1-\alpha} P(A) \text{ for all } A \in \mathbb{B}(\mathbb{R}) \right\}.$$

Definition 2.4. *The Wasserstein distance between two probability measures $P,\ Q \in \mathcal{P}$, is defined as*

$$\mathcal{W}_2^2(P, Q) = \inf\{ \int ||x - y||^2 \mu(dx, dy),\ \mu \in M(P, Q) \}, \tag{2.2}$$

where $M(P, Q)$ is the set of probability measures on $\mathbb{R} \times \mathbb{R}$ with marginals P and Q.

The Wasserstein distance is related to the transportation problem, it is the minimum cost of "transporting" the mass of P to Q. If the measures are the same, then the transportation cost is equal to zero. In the real line, it can be computed as

$$\mathcal{W}_2^2(P, Q) = \int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)| \mathrm{d}u,$$

with F_P^{-1} and F_Q^{-1} the quantile functions of P and Q.
Now, if we have two contaminated probability measures P and Q, we could trim a portion α from P and Q to make them more similar. If we continue trimming, we would like to know which is the "best" level of trimming, the level that makes P and Q equal.
From **Proposition 2** in Alvarez-Esteban et al. (2012), it follows that $\mathcal{W}_2^2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) > 0$ if and only if $d_{TV}(P, Q) > \alpha$, P and Q as in (2.1). So, the total variation distance is the minimal level of trimming required to make P and Q equal, $\mathcal{W}_2^2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$.

In the case of spectral densities, the estimation of the spectrum has noise. In terms of a model, the signals $X_1(t)$ and $X_2(t)$ are observed with uncorrelated noises $N_1(t)$ and $N_2(t)$, i.e., we observe

$$X_i^O(t) = X_i(t) + N_i(t), \qquad i = 1, 2.$$

Where the unobserved process (the true signal) is contaminated by the noise $N_i(t)$. This gives rise to the spectrum in the following sense,

$$f_i^O(\omega) = f^{X_i}(\omega) + f^{N_i}(\omega), \qquad i = 1, 2.$$

This causes the estimations to be different, even in the case of two processes with the same spectral density, $f^{X_1}(\omega) = f^{X_2}(\omega)$. So, the total variation distance quantifies the level of similarity between two spectra, in the sense of (2.1).

2.2 TV distance to compare spectra

Our principal goal is to use the TV distance as a similarity measure between spectral densities. Since a spectral density does not necessarily integrate 1, we must normalize them before we compute the TV distance. It means that the TV distance will be able to identify changes in the distribution of the energy and not necessarily changes in the total energy. However, the detection of changes based in the TV distance can be done together with or after another method that detects changes in the total energy, if required.

The TV distance between spectra is a metric on the following space. Let $\mathcal{M}$ be the set of equivalence classes defined on the space of spectral density functions, as follows: f_1 and f_2 are in the same equivalence class ($f_1 \sim f_2$)

if there exist a constant value $c > 0$ such that $f_1(\omega) = cf_2(\omega)$ for almost all ω. The non-negativity, symmetry and subadditivity properties are obtained by restricting the TV distance to this space. The identity of indiscernibles is satisfied on $\mathcal{M}$, since

$$d_{TV}(f,g) = 0 \qquad \Leftrightarrow \qquad \exists c > 0 \ f(\omega) = cg(\omega), \ \forall\omega \qquad \Leftrightarrow \qquad g, f \in [f].$$

There exist many important probability metrics that are used by statisticians and probabilists, Gibbs and Su (2002) present a summary of inequalities between them that could be useful in practice. One of these metrics, that is also bounded between [0,1] as the TV distance, is the Hellinger distance (d_H), which is defined as

$$d_H(f,g) = \left(\int (\sqrt{f} - \sqrt{g})^2 \right)^{1/2}.$$

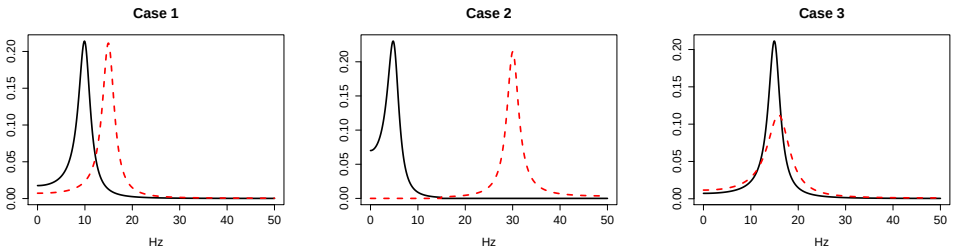
This metric is related to the TV distance through the inequalities

$$\frac{(d_H)^2}{2} \leq d_{TV} \leq d_H,$$

that could produce similar results when one compares two density functions. This distance is not considered in the rest of this thesis, however, it is still an option to explore. A possible disadvantage of the Hellinger distance is the lack of interpretation in terms of the spectral densities. In the frequency domain approach, the spectral density and the log spectral density have a physical interpretation while the interpretation of the square root of the density is not clear.

Chapter 1 described some of the distances used to compare spectral densities. Two of the most frequently employed are the L^2 norm and the Kullback-Leibler (KL) divergence. The KL is not symmetric, but there exists a symmetric version (SKL). Remember that, if f and g are two density functions,

$$\begin{aligned} d_{L^2}(f,g) &= \left(\int (f-g)^2 \right)^{1/2}, \\ d_{KL}(f,g) &= \int f \log\left(\frac{f}{g}\right) \\ \text{and} \quad d_{SKL}(f,g) &= (d_{KL}(f,g) + d_{KL}(g,f))/2. \end{aligned}$$



(a) (b) (c)

Figure 2.2: (a) Spectra with different peak frequency but close support. (b) Spectra with disjoint support. (c) Spectra with close peak frequency and similar support but different dispersion.

We illustrate with an example, some advantages of the TV distance over L^2 and SKL. Consider two unimodal spectra, as those presented in Figure 2.2. We look at three different cases. Case 1 - The first spectrum peaks at 10 Hz (black continuous curve) and the second peaks at 15 Hz. Case 2 - The first spectrum peaks at 5 Hz, the second peaks at 30 Hz, and the supports are disjoint. Case 3 - The first spectrum peaks at 15 Hz, the second peaks at 16 Hz, and they have different dispersion.

For each case, we compute the TV distance, the L^2 distance, and the SKL divergence. Table 2.1 shows these values. When the spectra are different as in Case 1, we would expect all distances to be "big", and indeed, all the considered distance values are big enough to distinguish between them. Now, if we observe Case 2, the spectra are completely different, since they have different supports. This example shows one of the disadvantages of the SKL divergence, since we cannot compute the value in this case. Notice that the TV distance has no problem, and the value is equal to one, which indicates that the spectra are completely different. On the other hand, the L^2 distance has a value comparable to that of Case 1, even though in this case the densities have disjoint support. Then, in Case 3 the spectra are different but it could be difficult to conclude that from the L^2 and SKL distances. The difference would be clearer using the TV distance. A more exhaustive simulation exercise to compare these distances is performed in Chapter 3.

The TV distance has a finite range, however, we need to establish a statistical notion of "big". This is important because, even in the case of two

Distance	Case 1	Case 2	Case 3
TV	0.686	1	0.232
L^2	0.402	0.5	0.146
SKL	1.413	NaN	0.141

Table 2.1: Distance values of the TV distance, L^2 norm and SKL divergence between the spectra plot in Figure 2.2.

samples with the same spectral representation, the estimated spectra have a TV distance value not equal to zero. We would like to choose a threshold for the TV distance between estimated spectra to decide if the samples were generated from the same spectral density or not, so that the probability of type I error is controlled at some level α. In addition, the procedure to choose this threshold must have enough power to detect when the true spectra are different. The next sections deal with the distribution of the TV distance between estimated spectral densities.

2.3 Distribution of the TV distance between estimated spectra

Let $X_1(t)$ and $X_2(t)$ be two time series with spectra f^{X_1}, f^{X_2}, and normalized spectra defined as

$$f_N^{X_i} = \frac{f^{X_i}}{\int_{1/2}^{1/2} f^{X_i}(\omega)\mathrm{d}\omega}, \qquad i = 1, 2.$$

Then, the TV distance between the normalized spectra will be

$$\begin{aligned} d_{TV}(f_N^{X_1}, f_N^{X_2}) &= 1 - \int_{-1/2}^{1/2} \min(f_N^{X_1}(\omega), f_N^{X_2}(\omega))\ d\omega \\ &= \frac{1}{2}\int_{-1/2}^{1/2} |f_N^{X_1}(\omega) - f_N^{X_2}(\omega)|\mathrm{d}\omega. \end{aligned} \tag{2.3}$$

2.3.1 Estimation of d_{TV}

At this point d_{TV}, defined as (2.3), is not a random quantity because it is based on the true (though unknown) spectral density. The next step is to

Window	Asymptotic Variance
Rectangular or Truncated	$\frac{2a}{T} f^2(\omega)$
Bartlett or Triangular	$\frac{2a}{3T} f^2(\omega)$
Daniell	$\frac{a}{T} f^2(\omega)$
Blackman - Tukey	$\frac{2a}{T}(1 - 4b + 6b^2) f^2(\omega)$
Parzen	$\frac{151}{280}\frac{a}{T} f^2(\omega)$

Table 2.2: Asymptotic variance for different windows. T is the length of the time series and a is the bandwidth.

consider a numeric approximation of d_{TV}. Then, we add uncertainty when f^{X_1} and f^{X_2} are unknown since we have to use estimations of them.
We apply the trapezoid method, to get a numerical approximation of equation (2.3). The trapezoid rule, which approximates the area under a curve using trapezoids instead of rectangles, is given by the following formula:

$$\int_b^c d(x)\mathrm{d}x \approx \frac{c-b}{n}\left(\frac{d(b)+d(c)}{2} + \sum_{k=1}^{n-1} d\left(b + \frac{k(c-b)}{n}\right)\right), \tag{2.4}$$

where n is the number of elements in the partition of the interval $[b, c]$. We could choose another numerical approximation and the procedure to obtain the asymptotic distribution would be similar.

For real data sets, f^{X_1} and f^{X_2} are not known and have to be estimated. As we mentioned before, the raw periodogram is not mean-square consistent because its variance does not decrease even when the length of the time series increases. So, we choose the lag window estimator (smoothed periodogram) defined in (1.2), with a Parzen window of width a. A lag window estimator can be rewritten as a spectral average estimator, i.e., the properties of the lag window estimators are similar to the spectral average.

The Parzen window is defined as

$$\beta(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & \text{if } |x| < \frac{1}{2} \\ 2(1-|x|)^3, & \text{if } \frac{1}{2} \leq |x| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

A possible criterion to choose a window is to compare the asymptotic variance of the resulting lag window estimator. Table 2.2 shows the asymptotic variance for different windows, the details can be found in Brockwell and Davis (2006). The Parzen window has smaller variance compared to the Rectangular, Bartlett, and Daniell. However, the Blackman - Tukey window with parameter b could have a smaller variance (depending on b) than the Parzen window. The Parzen window is included in many computational packages, in particular it is implemented in the function *spec.parzen* of the *HSMClust* Toolbox in R that we developed (see Appendix A).

Finally, to normalize the estimated spectra, we use $\int_{-1/2}^{1/2} \hat{f}(\omega)\mathrm{d}\omega = \hat{\gamma}(0)$. Our final estimator of the spectra is

$$\hat{f}_N^{X_i}(\omega) = \frac{\hat{f}^{X_i}(\omega)}{\hat{\gamma}^{X_i}(0)}. \tag{2.5}$$

Using the estimator (2.5) and the numerical approximation (2.4), we can write an estimator of the TV distance, $\hat{d}_{TV}$, as follows.

$$\hat{d}_{TV} = \frac{1}{2T} \sum_{k=1}^{T} \left| \hat{f}_N^{X_1}\left(\frac{k}{T} - \frac{1}{2}\right) - \hat{f}_N^{X_2}\left(\frac{k}{T} - \frac{1}{2}\right) \right|, \tag{2.6}$$

where T is the time series length.
Remarks. 1) To obtain equation (2.6) we take $b = -1/2$, $c = 1/2$ and $n = T$, also we use the symmetry of $\hat{f}_N^{X_i}$.
2) We assume that T is even, if T is odd we consider $n = 2\lfloor T/2 \rfloor$, so the frequencies correspond to $w_k = k/T$, the fundamental Fourier frequencies. If X_1 and X_2 have different lengths, but one a multiple of the other to ensure that we have enough common Fourier frequencies, we consider the smaller T.

2.3.2 Asymptotic distribution of $\hat{d}_{TV}$

We would like to find the asymptotic distribution of (2.6) under H_0, so we can establish a critical value for the following test,

$$H_0 : f^{X_1}(\omega) = f^{X_2}(\omega),\ \forall\omega \quad vs \quad H_A : \exists\omega \text{ such that } f^{X_1}(\omega) \neq f^{X_2}(\omega).$$

Under H_0, we can write $\hat{d}_{TV}$ as follows. Let $f(\omega) = f^{X_1}(\omega) = f^{X_2}(\omega)$,

$$\hat{d}_{TV} = \frac{1}{2T}\sum_{k=1}^{T} f(\omega_k)\,|D_{T,k}| = \sum_{k=1}^{T} c_{T,k}\,|D_{T,k}|\,, \tag{2.7}$$

where $c_{T,k} = \frac{f(\omega_k)}{2T}$, $\omega_k = \left(\frac{k}{T} - \frac{1}{2}\right)$ and $D_{T,k} = \dfrac{\hat{f}_N^{X_1}(\omega_k) - \hat{f}_N^{X_2}(\omega_k)}{f(\omega_k)}$.
To find the distribution of $\hat{d}_{TV}$, we use the asymptotic convergence of $\hat{f}_N^{X_i}$ and the following property: If $\{X_t\}$ is a time series with $\mathrm{Var}(X_t) = \sigma^2$ and $t = 1, 2 \ldots, T$ then

$$\hat{\gamma}(0) \xrightarrow{P} \sigma^2, \tag{2.8}$$

when $T \longrightarrow \infty$.
Additionally, we are going to use the following lemmas and the Central Limit Theorem for triangular arrays.

Theorem 2.1. Let $\{(X_{n,j}, 1 \leq j \leq n),\ n \geq 1\}$ be a triangular array of row-wise independent random variables with mean zero. Set, for $n \geq 1$, $S_n = \sum_{j=1}^{n} X_{n,j}$, $s_{n,j}^2 = \mathbb{E}[X_{n,j}^2]$, and $s_n^2 = \mathrm{Var}(S_n) = \sum_{j=1}^{n} s_{n,j}^2$. If the Lindeberg condition

$$\lim_{n\to\infty}\sum_{j=1}^{n}\frac{1}{s_n^2}\int_{|X_{n,j}|>\varepsilon} X_{n,j}^2 d\mathrm{P} = 0$$

is satisfied, then

$$\frac{S_n}{s_n} \longrightarrow N(0,1).$$

Lemma 2.1. If the Lyapunov Condition: there exists a $\delta > 0$ such that

$$\frac{1}{s_n^{2+\delta}}\sum_{j=1}^{n}\mathbb{E}|X_{n,j}|^{2+\delta} \xrightarrow{T\to\infty} 0,$$

is satisfied then the Lindeberg condition is satisfied.

The proof of the theorem and lemma can be found in Brockwell and Davis (2006), Chapter 6, or Gnedenko and Kolmogorov (1968).

We shall now consider two processes $X_1(t)$ and $X_2(t)$ that satisfy the following hypothesis.

Assumption $\mathcal{A}1$. *Suppose that $X_1(t)$ and $X_2(t)$ are independent stationary processes with mean μ, variance σ^2, absolutely summable covariance functions and the spectral densities f^{X_i} are continuous functions with the first three spectral moments finite.*

Lemma 2.2. Let $\hat{f}_L(\omega) = (2\pi)^{-1} \sum_{|h|\leq a} \beta(h/a)\hat{\gamma}(h)e^{-ih\omega}$, where β is a taper function, a is the bandwidth, T is the length of the time series X_t, $\hat{\gamma}(h) = \sum_{t=1}^{T-|h|}(X_t - \overline{X})(X_{t+|h|} - \overline{X})$ the sample covariance function, and $\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$. Under assumption $\mathcal{A}1$, if $a \to \infty$ when $T \to \infty$ and $a/T \to 0$, then

$$\sqrt{\frac{T}{a}}\left(\hat{f}_L(\omega) - f(\omega)\right) \xrightarrow{w} N\left(0, f^2(\omega)\int_{-1}^{1}\beta^2(u)\mathrm{d}u\right) \tag{2.9}$$

The proof can be found in Brillinger (1981), Section 5.6.

Before stating the normal approximation theorem we introduce some notation:

$$\begin{aligned} \Sigma^2 &= \frac{2}{\sigma^4}\int_{-1}^{1}\beta^2(u)\mathrm{d}u, \qquad \mu_{T,k} = c_{T,k}\Sigma\sqrt{\frac{2a}{\pi T}} = f(\omega_k)\Sigma\sqrt{\frac{a}{2\pi T^3}}, \\ s_{T,k}^2 &= c_{T,k}^2\Sigma^2\left(1 - \frac{2}{\pi}\right)\frac{a}{T} \qquad \text{and} \qquad s_T^2 = \sum_{k=1}^{T} s_{T,k}^2. \end{aligned}$$

Theorem 2.2 (Normal Approximation)**.** Suppose assumption $\mathcal{A}1$ holds, then, under H_0

$$\frac{1}{s_T}\sum_{k=1}^{T}\left(c_{T,k}|D_{T,k}| - \mu_{T,k}\right) \xrightarrow{w} N(0,1) \tag{2.10}$$

when $T \to \infty$ and $a/T \to 0$

Proof. Let $\hat{f}_N^{X_i}$ be the normalized lag window estimator using the time series $X_i(t)$ with bandwidth a, T the length of the observed time series, $\sigma^2 = \mathrm{Var}(X_i(t))$ $i = 1, 2$, and $\hat{d}_{TV}$ as (2.6).
Let $Z_{T,k}$ and $Z^*_{T,k}$, $k = 1, \ldots, T$, be independent Gaussian random variables with parameters

$$\mu = \frac{1}{\sigma^2} \text{ and } \sigma^2_{a,T} = \frac{a}{T\sigma^4} \int_{-1}^{1} \beta^2(u) \mathrm{d}u.$$

Step 1. First, we will show that $\hat{d}_{TV} - \sum_{k=1}^{T} c_{T,k} |Z_{T,k} - Z^*_{T,k}|$ converges in probability to zero.

Notice that $c_{T,k} \geq 0$. Without loss of generality, we can assume that $c_{T,k} > 0$ (if not, the elements being summed up will be zero). Now, under H_0,

$$\begin{aligned}
\left| \hat{d}_{TV} - \sum_{k=1}^{T} c_{T,k} |Z_{T,k} - Z^*_{T,k}| \right| &= \left| \sum_{k=1}^{T} c_{T,k} \left(|D_{T,k}| - |Z_{T,k} - Z^*_{T,k}| \right) \right| \\
&\leq \sum_{k=1}^{T} c_{T,k} \big| \; |D_{T,k}| - |Z_{T,k} - Z^*_{T,k}| \; \big| \\
&\leq \sum_{k=1}^{T} c_{T,k} \big| \; D_{T,k} - (Z_{T,k} - Z^*_{T,k}) \; \big|
\end{aligned}$$

Then,

$$0 \; \leq \; \left| \hat{d}_{TV} - \sum_{k=1}^{T} c_{T,k} |Z_{T,k} - Z^*_{T,k}| \right| \; \leq \; \sum_{k=1}^{T} c_{T,k} \big| \; D_{T,k} - (Z_{T,k} - Z^*_{T,k}) \; \big|. \tag{2.11}$$

It is sufficient to show that the sum on the right in (2.11) converges in

probability to zero. Let $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{k=1}^{T} c_{T,k} \big|\ D_{T,k} - (Z_{T,k} - Z^*_{T,k})\ \big| > \varepsilon \Big) \\
&\qquad \leq \sum_{k=1}^{T} \mathbb{P}\left(c_{T,k} \big|\ D_{T,k} - (Z_{T,k} - Z^*_{T,k})\ \big| > \varepsilon \right) \\
&\qquad = \sum_{k=1}^{T} \mathbb{P}\left(\big|\ D_{T,k} - (Z_{T,k} - Z^*_{T,k})\ \big| > \frac{\varepsilon}{c_{T,k}} \right) \\
&\qquad = \sum_{k=1}^{T} \mathbb{P}\left(\big|\ \frac{\hat{f}^{X_1}_N(\omega_k) - \hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - (Z_{T,k} - Z^*_{T,k})\ \big| > \frac{\varepsilon}{c_{T,k}} \right) \\
&\qquad = \sum_{k=1}^{T} \mathbb{P}\left(\big|\ \frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} - \left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right)\ \big| > \frac{\varepsilon}{c_{T,k}} \right).
\end{aligned}$$

Now, for each k, applying Chebychev's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\ \frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} - \left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right)\ \right| > \frac{\varepsilon}{c_{T,k}} \right) \\
&\qquad \leq \frac{c^2_{T,k}}{\varepsilon^2} \mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} - \left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right) \right)^2.
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} - \left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right) \right)^2 = \mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} \right)^2 \\
&-2\mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} \right) \mathbb{E}\left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right) + \mathbb{E}\left(\frac{\hat{f}^{X_2}_N(\omega_k)}{f(\omega_k)} - Z^*_{T,k} \right)^2.
\end{aligned} \tag{2.12}$$

Now, we compute each term,

$$\mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} \right)^2 = Var\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} \right) + \left[\mathbb{E}\left(\frac{\hat{f}^{X_1}_N(\omega_k)}{f(\omega_k)} - Z_{T,k} \right) \right]^2. \tag{2.13}$$

Combining (2.8) and (2.9) in Lemma 2.2,

$$\mathbb{E}\left(\frac{\hat{f}_N^{X_1}(\omega_k)}{f(\omega_k)}\right) \longrightarrow \frac{1}{\sigma^2} \qquad \text{and} \qquad \sqrt{\frac{T}{a}} Var\left(\frac{\hat{f}_N^{X_1}(\omega_k)}{f(\omega_k)}\right) \longrightarrow \frac{1}{\sigma^4}\int_{-1}^{1}\beta^2(u)\mathrm{d}u.$$

Since $Z_{T,k}$ is independent of X_1,

$$\begin{aligned} \mathbb{E}\left(\frac{\hat{f}_N^{X_1}(\omega_k)}{f(\omega_k)} - Z_{T,k}\right) &\longrightarrow 0, \text{ and} \\ \sqrt{\frac{T}{a}} Var\left(\frac{\hat{f}_N^{X_1}(\omega_k)}{f(\omega_k)} - Z_{T,k}\right) &\longrightarrow \frac{1}{\sigma^4}\int_{-1}^{1}\beta^2(u)\mathrm{d}u < \infty. \end{aligned} \tag{2.14}$$

The same proof works for X_2 and $Z^*_{T,k}$. Denote

$$G_{T,k} = \frac{\hat{f}_N^{X_1}(\omega_k)}{f(\omega_k)} - Z_{T,k}$$

and

$$G^*_{T,k} = \frac{\hat{f}_N^{X_2}(\omega_k)}{f(\omega_k)} - Z^*_{T,k}.$$

Substituting this notation and using the previous inequalities we get.

$$\begin{aligned} \sum_{k=1}^{T}\mathbb{P}\bigg(\left|G_{T,k} - G^*_{T,k}\right| > \frac{\varepsilon}{c_{T,k}}\bigg) &\leq \sum_{k=1}^{T}\frac{c^2_{T,k}}{\varepsilon^2}\mathbb{E}\left(G_{T,k} - G^*_{T,k}\right)^2 \\ &= \sum_{k=1}^{T}\frac{c^2_{T,k}}{\varepsilon^2}\Big(\mathbb{E}[G^2_{T,k}] - 2\mathbb{E}[G_{T,k}]\mathbb{E}[G^*_{T,k}] + \mathbb{E}[(G^*_{T,k})^2]\Big) \\ &= \sum_{k=1}^{T}\frac{c^2_{T,k}}{\varepsilon^2}\Big(\mathrm{Var}[G_{T,k}] + \mathbb{E}^2[G_{T,k}] - 2\mathbb{E}[G_{T,k}]\mathbb{E}[G^*_{T,k}] \\ &\qquad + \mathrm{Var}[G^*_{T,k}] + \mathbb{E}^2[G^*_{T,k}]\Big). \end{aligned}$$

Notice that the moments of $G_{T,k}$ and $G^*_{T,k}$ are equal and have the same value

for all k. So, if k_0 denotes a fixed value of k,

$$\begin{aligned}\sum_{k=1}^{T}\mathbb{P}\left(\left|G_{T,k}-G^*_{T,k}\right|>\frac{\varepsilon}{c_{T,k}}\right) &\leq \left(\sum_{k=1}^{T}\frac{c^2_{T,k}}{\varepsilon^2}\right)2\mathrm{Var}[G_{T,k_0}]\\ &= \left(\sum_{k=1}^{T}\frac{f^2(\omega_k)}{T^2\varepsilon^2}\right)2\left(\sqrt{\frac{a}{T}}\left(\sqrt{\frac{T}{a}}\mathrm{Var}[G_{T,k_0}]\right)\right)\\ &= \frac{2}{T\varepsilon^2}\left(\frac{1}{T}\sum_{k=1}^{T}f^2(\omega_k)\right)\sqrt{\frac{a}{T}}\left(\sqrt{\frac{T}{a}}\mathrm{Var}[G_{T,k_0}]\right)\\ &= \frac{2a^{1/2}}{T^{3/2}\varepsilon^2}\left(\frac{1}{T}\sum_{k=1}^{T}f^2(\omega_k)\right)\left(\sqrt{\frac{T}{a}}\mathrm{Var}[G_{T,k_0}]\right)\end{aligned}$$

Assuming that the first three spectral moments are finite then

$$\frac{1}{T}\sum_{k=1}^{T}f^2(\omega_k)\overset{T\to\infty}{\longrightarrow}\int_{-1/2}^{1/2}f^2(\omega)\mathrm{d}\omega. \tag{2.15}$$

Using (2.15) and (2.14) we get that

$$\sum_{k=1}^{T}\mathbb{P}\left(\left|G_{T,k}-G^*_{T,k}\right|>\frac{\varepsilon}{c_{T,k}}\right)\longrightarrow 0, \tag{2.16}$$

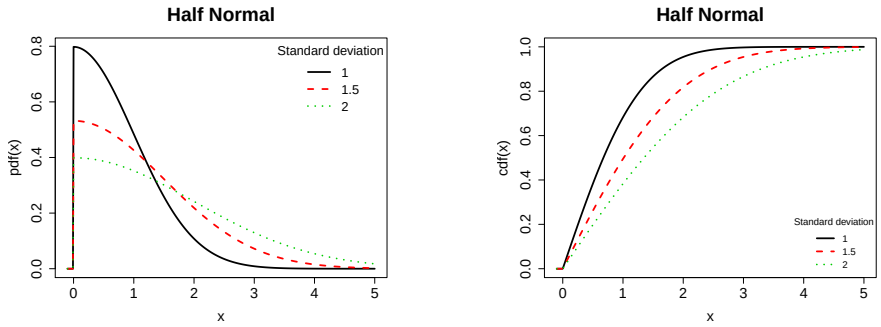
when $T\to\infty$. The convergence of (2.16) and the bound in (2.11) prove that $\hat{d}_{TV}-\sum_{k=1}^{T}c_{T,k}|Z_{T,k}-Z^*_{T,k}|$ converges in probability to zero.

Step 2. Now, we show that

$$\frac{1}{s_T}\sum_{k=1}^{T}\left(c_{T,k}|Z_{T,k}-Z^*_{T,k}|-\mu_{T,k}\right)\overset{w}{\longrightarrow}N(0,1).$$

We have a triangular array, so we would like to use the **Theorem 2.1**. Let T and k be fixed, then

$$\begin{aligned}Z_{T,k}-Z^*_{T,k} &\sim N\left(0,\frac{a}{T}\Sigma^2\right),\\ |Z_{T,k}-Z^*_{T,k}| &\sim HN\left(\frac{a}{T}\Sigma^2\right),\end{aligned} \tag{2.17}$$



(a) Density function (b) Distribution function

Figure 2.3: Probability density and distribution functions of the Half Normal (HN) distribution with different standard deviation values.

where $\Sigma^2 = \frac{2}{\sigma^4}\int_{-1}^{1}\beta^2(u)\mathrm{d}u$ and HN denotes the Half-Normal distribution. The HN distribution is a particular case of the folded normal (see Leone et al., 1961), when the mean is equal to zero. This distribution is used when the measurements are Gaussian but only their absolute value is considered. Figure 2.3 plots examples of the density and distribution functions.

Using properties of this distribution,

$$\mathbb{E}(|Z_{T,k} - Z^*_{T,k}|) = \Sigma\sqrt{\frac{2a}{\pi T}} \tag{2.18}$$

$$\mathrm{Var}(|Z_{T,k} - Z^*_{T,k}|) = \Sigma^2\left(1 - \frac{2}{\pi}\right)\frac{a}{T}. \tag{2.19}$$

Let $Y_{T,k} = c_{T,k}|Z_{T,k} - Z^*_{T,k}| - \mu_{T,k}$, where

$$\mu_{T,k} = c_{T,k}\Sigma\sqrt{\frac{2a}{\pi T}} = m_1 f(\omega_k)\sqrt{\frac{a}{T^3}},$$

and m_1 is a constant. Then, $\{(Y_{T,k}, 1 \leq k \leq T),\ k \geq 1\}$ is an independent triangular array with mean zero and variance

$$s^2_{T,k} = c^2_{T,k}\Sigma^2\left(1 - \frac{2}{\pi}\right)\frac{a}{T} = m_2 f^2(\omega_k)\frac{a}{T^3},$$

where m_2 is a constant. Now, we need to verify the Lindeberg condition. In

fact, we will verify the Lyapunov condition, since it implies the Lindeberg condition.
Lyapunov Condition. There exist a $\delta > 0$ such that

$$\frac{1}{s_T^{2+\delta}} \sum_{k=1}^{T} \mathbb{E}|Y_{T,k}|^{2+\delta} \overset{T\to\infty}{\longrightarrow} 0.$$

Let $\delta = 1$, then using the density of the Half Normal distribution,

$$\begin{aligned}
\mathbb{E}|Y_{T,k}|^3 &= c_{T,k}^3 \int_0^\infty \left| u - \Sigma\sqrt{\frac{2a}{\pi T}} \right|^3 \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-\frac{u^2}{2\Sigma\sqrt{\frac{a}{T}}}} \mathrm{d}u \\
&= c_{T,k}^3 \int_0^{\Sigma\sqrt{\frac{2a}{\pi T}}} \left(u - \Sigma\sqrt{\frac{2a}{\pi T}} \right)^3 \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-\frac{u^2}{2\Sigma\sqrt{\frac{a}{T}}}} \mathrm{d}u \\
&- c_{T,k}^3 \int_{\Sigma\sqrt{\frac{2a}{\pi T}}}^\infty \left(y - \Sigma\sqrt{\frac{2a}{\pi T}} \right)^3 \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-\frac{u^2}{2\Sigma\sqrt{\frac{a}{T}}}} \mathrm{d}u. \qquad (2.20)
\end{aligned}$$

To compute (2.20), we use integration by parts and properties of the exponential function. Since we are interested in the limit, we will denote by m_i with $i = 3, 4, 5$, terms that do not depend of a, T or k. Then

$$\begin{aligned}
\mathbb{E}|Y_{T,k}|^3 &= c_{k,T}^3 \left[m_3 \left(\frac{a}{T}\right)^{3/2} + m_4 \left(\frac{a}{T}\right)^{3/2} \Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}} \right) \right] \\
&= \frac{1}{T^3} f^3(\omega_k) \left[m_3 \left(\frac{a}{T}\right)^{3/2} + m_4 \left(\frac{a}{T}\right)^{3/2} \Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}} \right) \right] \\
&= \frac{1}{T} f^3(\omega_k) \left[m_3 \left(\frac{a^{3/2}}{T^{7/2}}\right) + m_4 \left(\frac{a^{3/2}}{T^{7/2}}\right) \Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}} \right) \right] \qquad (2.21)
\end{aligned}$$

Then,

$$s_T^3 = m_5 \, \frac{a^{3/2}}{T^{9/2}} \left(\sum_{k=1}^{T} f^2(\omega_k) \right)^{3/2} = m_5 \, \frac{a^{3/2}}{T^3} \left(\frac{1}{T} \sum_{k=1}^{T} f^2(\omega_k) \right)^{3/2}.$$

So,

$$\begin{aligned}\frac{1}{s_T^3}\sum_{k=1}^{T}\mathbb{E}|Y_{T,k}|^3 &= \frac{\frac{1}{T}\sum_{k=1}^{T} f^3(\omega_k)\left[m_3\left(\frac{a^{3/2}}{T^{7/2}}\right)+m_4\left(\frac{a^{3/2}}{T^{7/2}}\right)\Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}}\right)\right]}{m_5\ \frac{a^{3/2}}{T^3}\left(\frac{1}{T}\sum_{k=1}^{T} f^2(\omega_k)\right)^{3/2}} \\ &= \frac{\frac{1}{T}\sum_{k=1}^{T} f^3(\omega_k)\left[m_3\left(\frac{1}{T^{1/2}}\right)+m_4\left(\frac{1}{T^{1/2}}\right)\Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}}\right)\right]}{m_5\left(\frac{1}{T}\sum_{k=1}^{T} f^2(\omega_k)\right)^{3/2}}.\end{aligned} \tag{2.22}$$

Notice that,

$$\begin{aligned}\Phi\left(\Sigma\sqrt{\frac{2}{\pi}}\sqrt{\frac{a}{T}}\right) &\overset{T\to\infty}{\longrightarrow} 1/2, \\ m_3\left(\frac{1}{T^{1/2}}\right) &\overset{T\to\infty}{\longrightarrow} 0, \\ m_4\left(\frac{1}{T^{1/2}}\right) &\overset{T\to\infty}{\longrightarrow} 0,\end{aligned}$$

then,

$$m_3\left(\frac{1}{T^{1/2}}\right)+m_4\left(\frac{1}{T^{1/2}}\right)\Phi\left(\Sigma\sqrt{\frac{2a}{\pi T}}\right)\overset{T\to\infty}{\longrightarrow} 0.$$

Since $\omega_k = (k/T - 1/2)$ and assuming that the first three spectral moments are finite then

$$\begin{aligned}\frac{1}{T}\sum_{k=1}^{T} f^3(\omega_k) &\overset{T\to\infty}{\longrightarrow} \int_{-1/2}^{1/2} f^3(\omega)\mathrm{d}\omega, \\ \frac{1}{T}\sum_{k=1}^{T} f^2(\omega_k) &\overset{T\to\infty}{\longrightarrow} \int_{-1/2}^{1/2} f^2(\omega)\mathrm{d}\omega.\end{aligned}$$

Finally, we get the Lyapunov condition,

$$\frac{1}{s_T^3}\sum_{k=1}^{T}\mathbb{E}|Y_{T,k}|^3 \overset{T\to\infty}{\longrightarrow} 0.$$

So,

$$\frac{1}{s_T}\sum_{k=1}^{T} Y_{T,k} = \frac{1}{s_T}\left(\sum_{k=1}^{T} c_{T,k}|Z_{T,k} - Z^*_{T,k}| - \mu_{T,k}\right) \xrightarrow{w} N(0,1),$$

when $T \to \infty$.

Finally, we conclude from **Step 1** and **Step 2** that $\hat{d}_{TV}$ converges to the same distribution as $\sum_{k=1}^{T} c_{T,k}|Z_{T,k} - Z^*_{T,k}|$. It means that $\hat{d}_{TV}$ is asymptotically Normal with the same parameters as $\sum_{k=1}^{T} c_{T,k}|Z_{T,k} - Z^*_{T,k}|$.

Remark. We could use directly the distribution of $\xi_{T,k} = |Z_{T,k} - Z^*_{T,k}|$ instead of two Gaussian variables, however to make the second step more intuitive we prefer to use two Gaussian variables.

2.3.3 Approximation of the distribution of $\hat{d}_{TV}$

Using **Theorem 2.2** we get an asymptotic distribution for $\hat{d}_{TV}$, however for finite T we need an alternative to approximate this distribution. From the proof of the theorem, we will use as an approximation the distribution of

$$\sum_{k=1}^{T} c_{T,k}|Z_{T,k} - Z^*_{T,k}|,$$

where $Z_{T,k}$, $Z^*_{T,k}$ are independent Gaussian variables as before.

A similar approximation can be obtained but considering chi-squared variables. Using the chi-squared approximation of the smoothed periodogram,

$$\frac{\hat{f}^{X_i}(\omega_k)}{f(\omega_k)} \overset{\cdot}{\sim} \frac{\chi^2_{2L_h}}{2L_h}, \tag{2.23}$$

where $L_h = \dfrac{T}{a\int_{-1}^{1}\beta^2(u)\mathrm{d}u}$.
Then the chi-squared approximation is

$$\sum_{k=1}^{T} c_{T,k}|Z_{T,k} - Z^*_{T,k}|,$$

where $c_{T,k} = \frac{f(\omega_k)}{2T}$, $Z_{T,k}$ and $Z^*_{T,k}$ $k = 1, \ldots, T$ are independent random variables with distribution,

$$\sqrt{\frac{\sigma^2}{2\pi}} Z^*_k \sim \frac{\chi^2_{2L_h}}{2L_h}, \text{ with } L_h = \frac{T}{a\int_{-1}^{1} \beta^2(u)\mathrm{d}u}.$$

2.3.4 Bootstrapping

As an alternative to the asymptotic distribution, one can obtain a critical value of the statistic $\hat{d}_{TV}$ based on a bootstrap procedure. If $X(t)$ is a linear process, i.e.

$$X(t) = \sum_{j=-\infty}^{\infty} \psi_j W(t-j)$$

where $W(t)$ is white noise, it can be proved (see Bloomfield, 1976) that the periodogram of $X(t)$ satisfies

$$I^X(\omega_k) \approx G(\psi, \omega) I^W(\omega_k),$$

where $G(\psi, \omega) = \sum \psi_j e^{-2\pi i \omega j}$ and $I^W(\cdot)$ is the periodogram of the white noise $W(t)$. Moreover, if the white noise has variance equal to one then $G(\psi, \omega)$ is equal to the spectral density of X. So, the observed periodogram consist of the spectral density of X multiplied by the periodogram of a white noise process. A natural way to get a replicate of the observed spectral density is to multiply the density f times the estimated periodogram of white noise. We will explain this with more detail when the consistency of the bootstrap estimator is proved.
This proposal is motivated by the method presented in Kreiss and Paparoditis (2015).

Algorithm:

1. From $X_1(t)$ and $X_2(t)$, estimate $\hat{f}_N^{X_1}(\omega)$ and $\hat{f}_N^{X_2}(\omega)$.
2. Under H_0, take $\hat{f}_N(\omega) = \dfrac{\hat{f}_N^{X_1}(\omega) + \hat{f}_N^{X_2}(\omega)}{2}$.
3. Draw $Z(1), \ldots, Z(T) \sim N(0,1)$ i.i.d random variables, then estimate $\hat{f}_N^{Z}(\omega)$ using also the lag window estimator.

4. The bootstrap spectral density will be

$$\hat{f}_N^B(\omega) = \hat{f}_N(\omega)\hat{f}_N^Z(\omega).$$

5. Repeat 3 and 4 and estimate $\hat{d}_{TV}$ using the bootstrap spectral densities, i.e.,

$$\hat{d}_{TV}^B = \hat{d}_{TV}(\hat{f}_N^{B_1}, \hat{f}_N^{B_2}),$$

where $\hat{f}_N^{B_i}$, $i = 1, 2$, are two bootstrap spectral densities using different replicates of the process $Z(\cdot)$.

In order to have a good approximation of the distribution of $\hat{d}_{TV}$ by $\hat{d}_{TV}^B$, we need to show that $\hat{f}_N^B(\omega)$ is a consistent estimator.

Consistency of the Bootstrap estimator for the spectral density. Suppose that $X(t)$ is a linear processes, i.e. $X(t) = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ where Z_t is white noise with variance equal to 1, and assume that $\sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty$ and $\mathbb{E}Z_1^4 < \infty$. From **Theorem 10.3.1** in Brockwell and Davis (2006), we know that,

$$I^X(\omega_k) = |\psi(e^{-i\omega_k})|^2 I^Z(\omega_k) + R_T(\omega_k), \tag{2.24}$$

where $\psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda j}$ and $\max_{\omega_k \in [0,\pi]} \mathbb{E}|R_T(\omega_k)|^2 = O(T^{-1})$. In addition, X is a linear filter of Z, so

$$f^X(\omega) = |\psi(e^{-i\omega})|^2 f^Z(\omega).$$

Because $Z(t)$ is white noise with variance one, $f^Z(\omega) = 1$. Then,

$$f^X(\omega) = |\psi(e^{-i\omega})|^2. \tag{2.25}$$

Substituting (2.25) into (2.24), we obtain

$$I^X(\omega_k) = f^X(\omega_k) I^Z(\omega_k) + R_T(\omega_k). \tag{2.26}$$

Define the functional J as $J(f)(\omega) = f_N^X(\omega)\hat{f}_N^Z(\omega)$, where $\hat{f}^Z$ is the lag window estimator for the spectral density of Gaussian standard white noise. Since we are interested in bootstrapping the lag window estimator instead of just the periodogram, we now study the behavior of $J(f)(\omega)$. We shall consider the equivalent representation of a lag window estimator as an averaged periodogram. If

$$\hat{f}_L(\omega) = \sum_{|h|\leq a} \beta(h/a)\hat{\gamma}(h)e^{-i2\pi h\omega},$$

is the lag window estimator with bandwidth a, then we can approximate $\hat{f}_L$ as

$$\hat{f}_L(\omega) \approx \sum_{|j|<\lfloor (T-1)/2 \rfloor} \boldsymbol{\beta}_T(\omega_j)I(\omega_j),$$

where $\boldsymbol{\beta}_T(\omega) = \frac{2\pi}{T}\sum_{|h|\leq a}\beta(h/a)e^{-i2\pi h\omega}$ and $I(\omega_j)$ is the periodogram (see Brockwell and Davis, 2006). To prove the consistency of the bootstrap, it is necessary to assume that $\boldsymbol{\beta}$ decreases exponentially and to consider that f_N^X is smooth. This is valid when one uses the Parzen window, but it is not necessarily true in other cases.
Then,

$$\begin{aligned} J(f)(\omega) = f_N^X(\omega_k)\hat{f}_N^Z(\omega_k) &\approx f_N^X(\omega_k)\sum_j \boldsymbol{\beta}_T(\omega_j)I^Z(\omega_{k+j}) \\ &= \sum_j \boldsymbol{\beta}_T(\omega_j)f_N^X(\omega_k)I^Z(\omega_{k+j}) \\ &\approx \sum_j \boldsymbol{\beta}_T(\omega_j)f_N^X(\omega_{k+j})I^Z(\omega_{k+j}) \qquad (2.27) \\ &= \sum_j \boldsymbol{\beta}_T(\omega_j)(I^X(\omega_{k+j}) - R_T(\omega_{k+j})) \qquad (2.28) \\ &= \sum_j \boldsymbol{\beta}_T(\omega_j)I^X(\omega_{k+j}) - \sum_j \boldsymbol{\beta}_T(\omega_j)R_T(\omega_{k+j}). \end{aligned}$$

So, applying (2.26), we can rewrite (2.27) as (2.28), and

$$J(f) = \hat{f}_N^X(\omega_k) - \tilde{R}_T(\omega_k),$$

where $\tilde{R}_T(\omega_k) = \sum_j \boldsymbol{\beta}_T(\omega_j)R_T(\omega_{k+j})$.
To prove consistency of the bootstrap procedure, it is enough to show that

$$\max_{\omega_k\in[0,\pi]} |J(f)(\omega_k) - \hat{f}_N^X(\omega_k)| \xrightarrow{P} 0. \qquad (2.29)$$

Considering the previous equations, we have

$$|J(f) - \hat{f}_N^X(\omega)| = |\tilde{R}_T(\omega_k)|.$$

From (2.24) we have

$$\begin{aligned}
\max_{\omega_k \in [0,\pi]} \mathbb{E}|R_T(\omega_k)|^2 = O(T^{-1}) &\Rightarrow \max_{\omega_k \in [0,\pi]} \mathbb{E}|R_T(\omega_k)|^2 = o(1) \\
&\Rightarrow \mathbb{E}|R_T(\omega_k)|^2 = o(1), \qquad \forall \omega_k \\
&\Rightarrow R_T(\omega_k) \xrightarrow{P} 0, \qquad \forall \omega_k \\
&\Rightarrow \max_{\omega_k \in [0,\pi]} |R_T(\omega_k)| \xrightarrow{P} 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\max_{\omega_k \in [0,\pi]} |\tilde{R}_T(\omega_k)| &= \max_{\omega_k \in [0,\pi]} \left| \sum_j \boldsymbol{\beta}(\omega_j) R_T(\omega_{k+j}) \right| \\
&\leq \sum_j \boldsymbol{\beta}(\omega_j) \max_{\omega_k \in [0,\pi]} |R_T(\omega_{k+j})| \xrightarrow{P} 0. \qquad (2.30)
\end{aligned}$$

We conclude from (2.30) that (2.29) holds and the bootstrap estimator of the spectral density is consistent. Then, the consistency of the bootstrap approximation for the distribution of $\hat{d}_{TV}$ is a consequence, since $\hat{d}_{TV}$ is a continuous function of the spectral density.

2.4 Simulation Study

The general setting of the simulation study is based on comparing (under H_0) the empirical distributions between 1) values of $\hat{d}_{TV}$, obtained by simulating time series $X_1(t)$ and $X_2(t)$ and computing the TV distance between the normalized estimated spectra, and 2) values of the Normal, Chi-square and/or Bootstrap approximations obtained by drawing from the variables Z_k and Z_k^*.
We consider two cases for $X_i(t),\ i = 1, 2$. Case 1 - AR(1) process with $\phi = .5$, and Case 2 - an AR(2) process with $(\phi_1, \phi_2) = (-.5, -.6)$. Figure 2.4 shows the spectra for these processes.

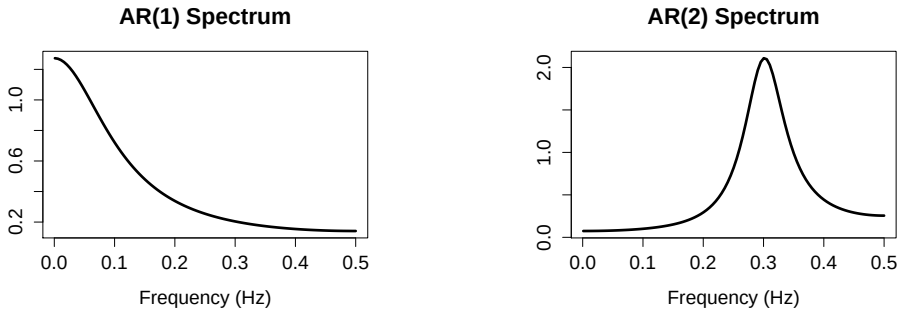


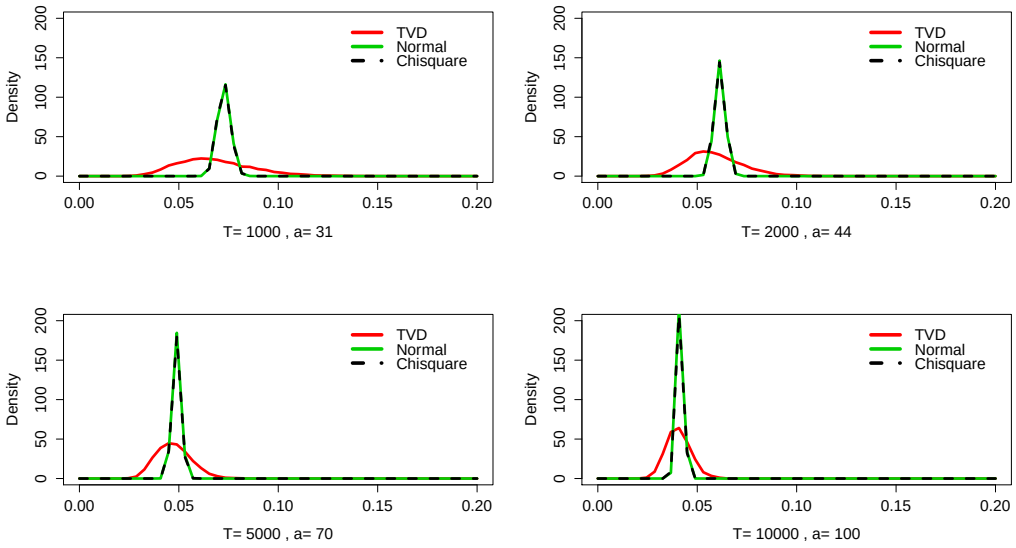
Figure 2.4: Spectra used in the simulation study to draw $X_i(t),\ i = 1, 2..$

2.4.1 Rate of convergence

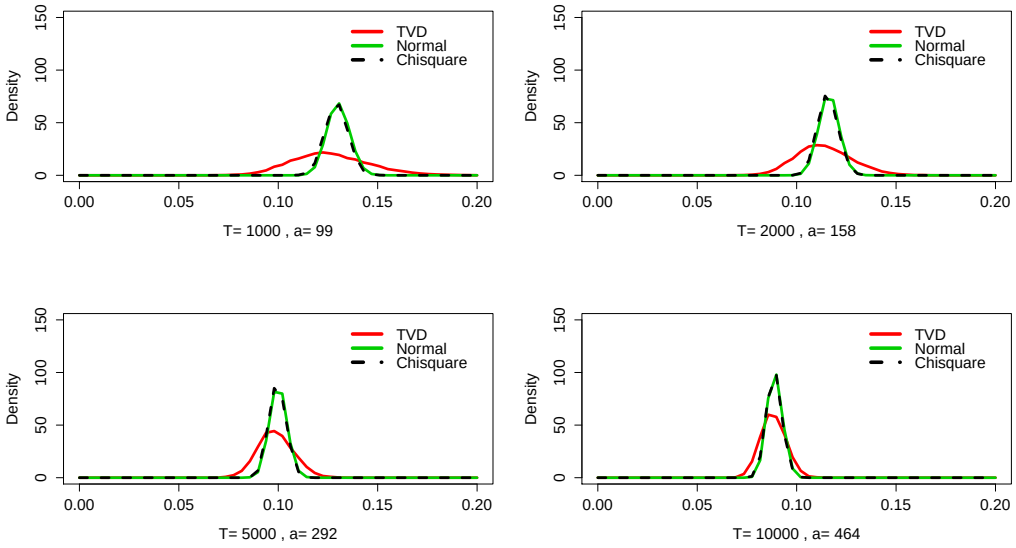
Notice that the asymptotic convergence requires that $\frac{a}{T} \to 0$. However, the theorem does not specify how to choose a. In the simulation study, we consider $a = T^{\frac{p-1}{p}}$ with $p = 2, 3, 4, 5$. Then $\frac{a}{T} = \frac{1}{T^{1/p}}$ and the condition is satisfied.
Other values for the simulation are $T = 1000, 2000, 5000$ and 10000 replicates to obtain the empirical distribution in each case. Our goal is to study the rate of convergence, but we are also interested in observing if the approximations are sensitive to the choice of a.

First, we shall focus on the Normal and Chi-square approximations, assuming the true spectral density known. Figures 2.5, 2.6, 2.7 and 2.8 show a nonparametric estimation of the density of each sample, the red density was estimated using the values of TV distance, the green density was estimated using the values of the normal approximation and the black dashed density was estimated using the values of the Chi-square approximation.

Although the spectra are very different, the results are similar in both examples. We can observe a good approximation in the cases $p \geq 3$ and when $p = 4$ we observe that the densities are almost the same for $T = 10000$. In the case $p = 2$ the convergence is slower and bigger values of T are needed in order to have a good approximation. In the other cases, a good approximation is observed for $T \geq 5000$.
It seems that the best choice for a could be $T^{4/5}$, "best" in the sense that the convergence will be faster.

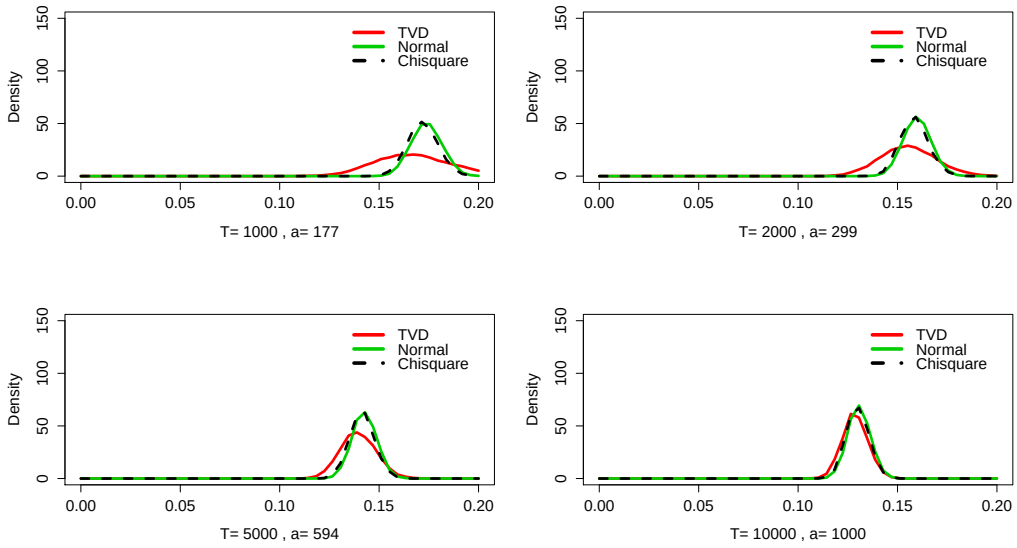


(a) $p = 2$

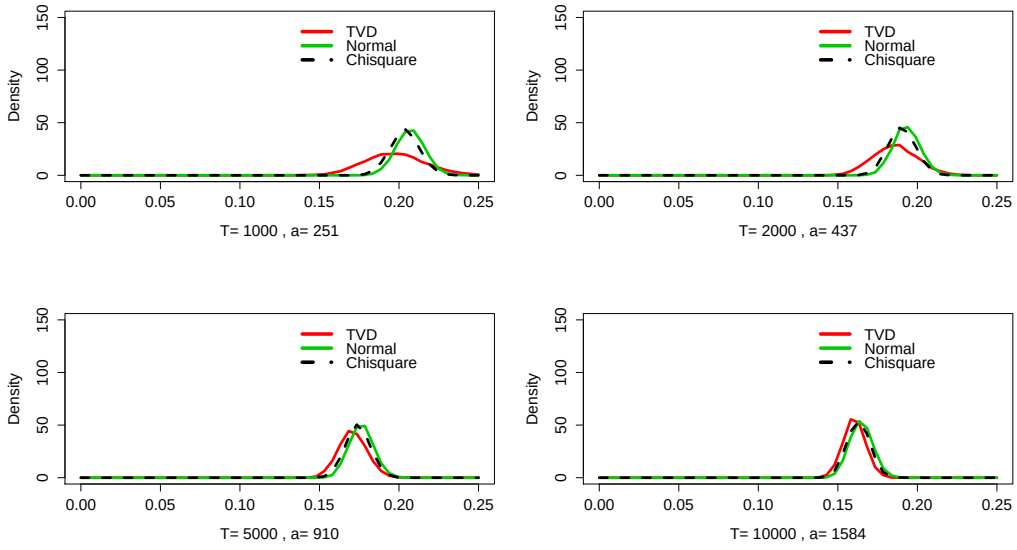


(b) $p = 3$

Figure 2.5: Simulation results for AR(1), $T = 1000, 2000, 5000, 10000$, $a = T^{\frac{p-1}{p}}$, $p = 2, 3$.

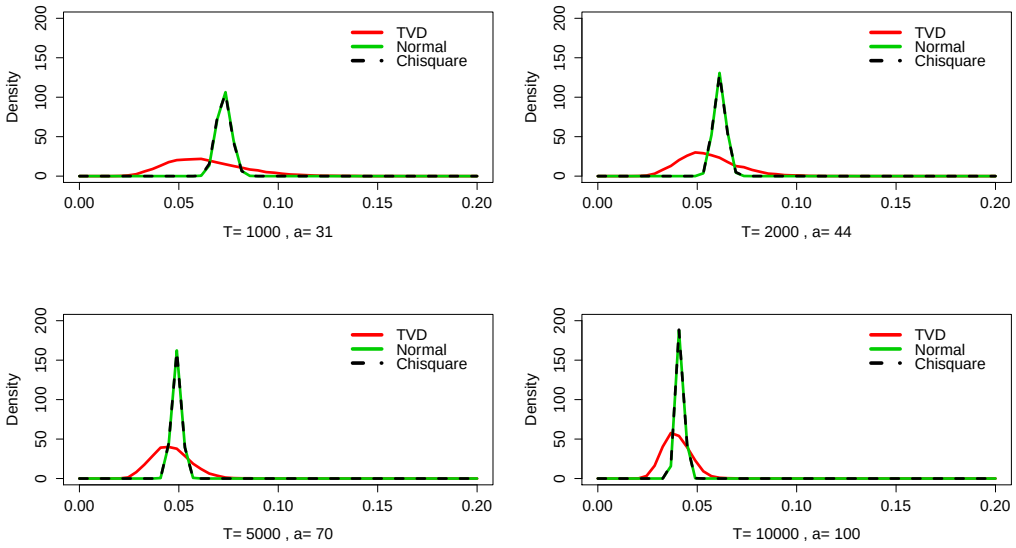


(a) $p = 4$

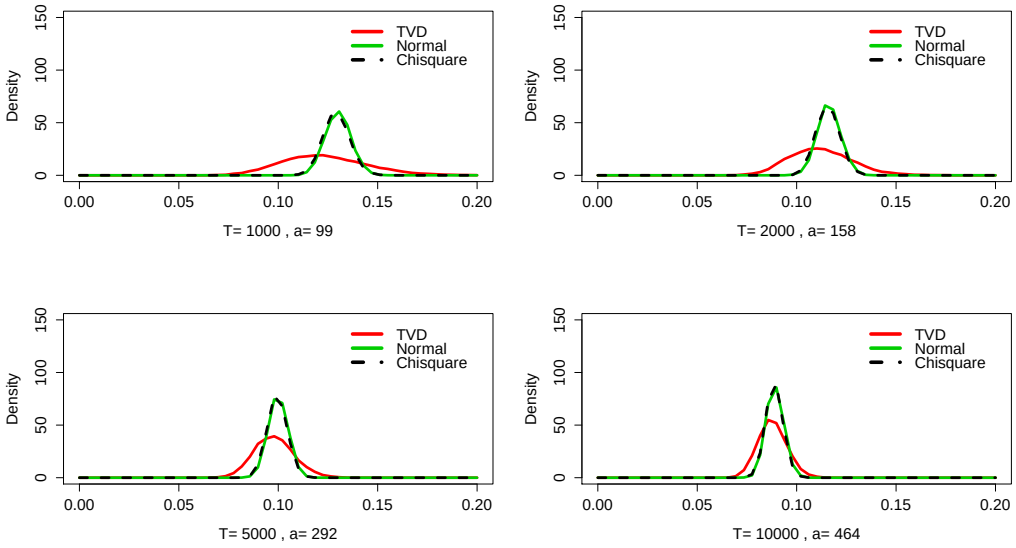


(b) $p = 5$

Figure 2.6: Simulation results for AR(1), $T = 1000, 2000, 5000, 10000$, $a = T^{\frac{p-1}{p}}$, $p = 4, 5$.

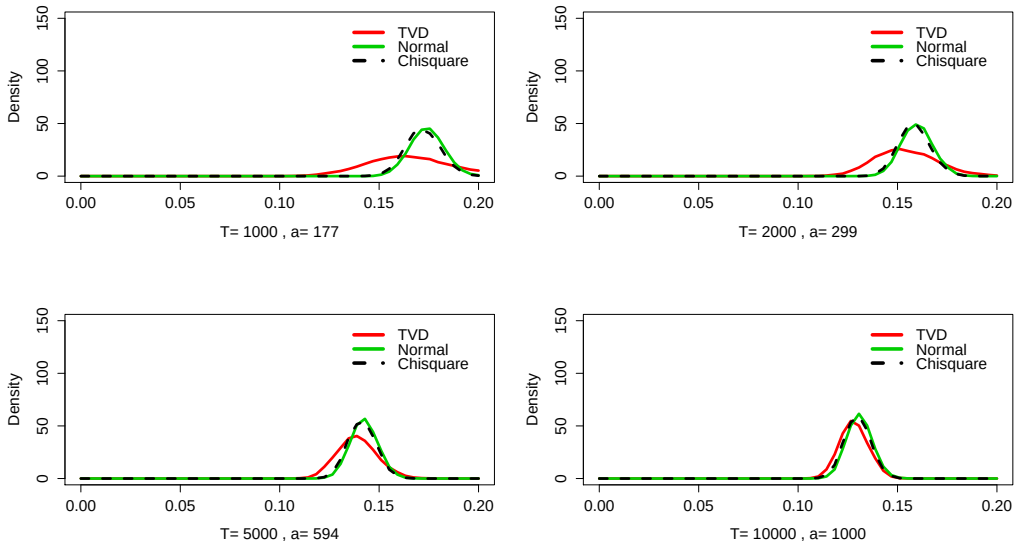


(a) $p = 2$

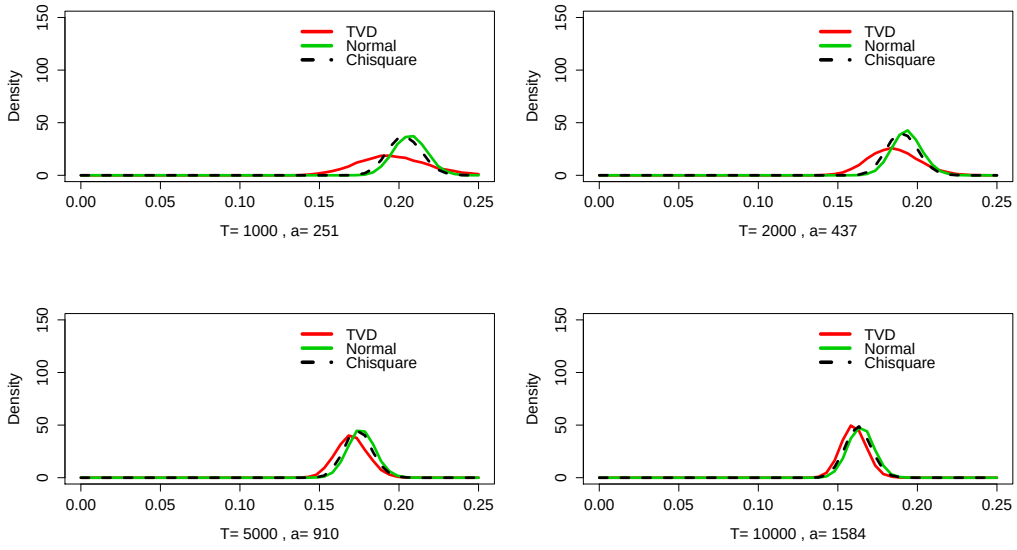


(b) $p = 3$

Figure 2.7: Simulation results for AR(2), $T = 1000, 2000, 5000, 10000$, $a = T^{\frac{p-1}{p}}$, $p = 2, 3$.



(a) $p = 4$



(b) $p = 5$

Figure 2.8: Simulation results for AR(2), $T = 1000, 2000, 5000, 10000$, $a = T^{\frac{p-1}{p}}$, $p = 4, 5$.

These results show that we should have at least 5000 points to get a good approximation. However, sometimes we are not able to get so much data. We explore the case of a "small" sample size and propose a modification to get a better approximation. "Small" will be relative to a time series case, because we cannot have a good estimation of the spectral density if T is too small. We consider the cases when $T = 1000$ and $T = 2000$.

First, we would like to verify if the convergence will be faster with a better choice of the bandwidth a. Consider the same AR(1) and AR(2) processes as before but with $a = 100, 200, 300$, and 400. Figures 2.9 and 2.10 show the results for each case. Increasing the value of a we can approximate better the dispersion of the $\hat{d}_{TV}$, however, a bias appears.

Why does this happen? Consider the Normal approximation, since Z_k and Z_k^* are not correlated, then

$$|Z_k - Z_k^*| \ \dot{\sim} \ HN\left(\tfrac{a}{T}\Sigma^2\right), \tag{2.31}$$

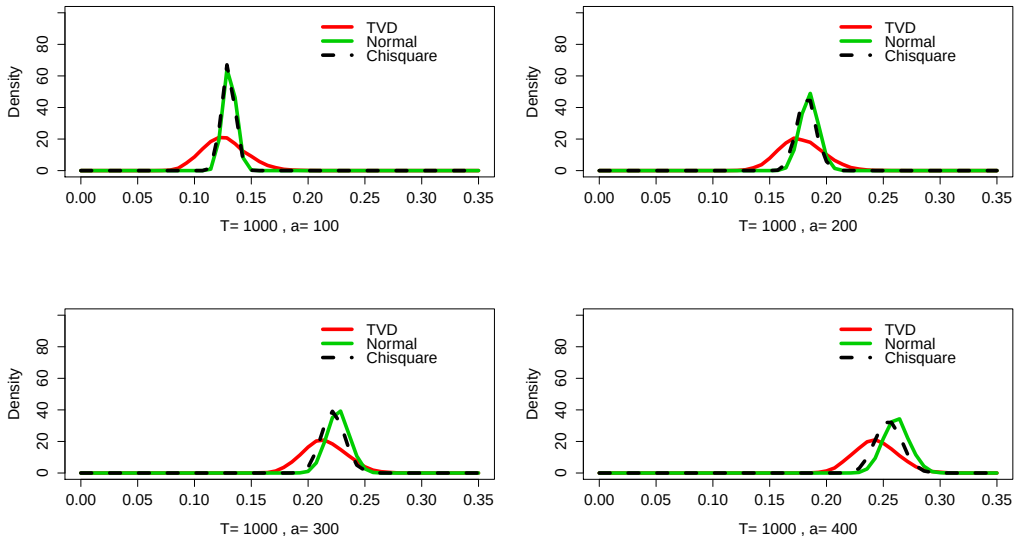
where $\Sigma^2 = \dfrac{2}{\sigma^4}\displaystyle\int_{-1}^{1} \beta^2(u)\mathrm{d}u$ and HN denotes the Half-Normal distribution. Using properties of this distribution,

$$\begin{aligned} \mathbb{E}(|Z_k - Z_k^*|) &= \Sigma\sqrt{\frac{2a}{\pi T}} \\ \mathrm{Var}(|Z_k - Z_k^*|) &= \Sigma^2\left(1 - \frac{2}{\pi}\right)\frac{a}{T}. \end{aligned}$$

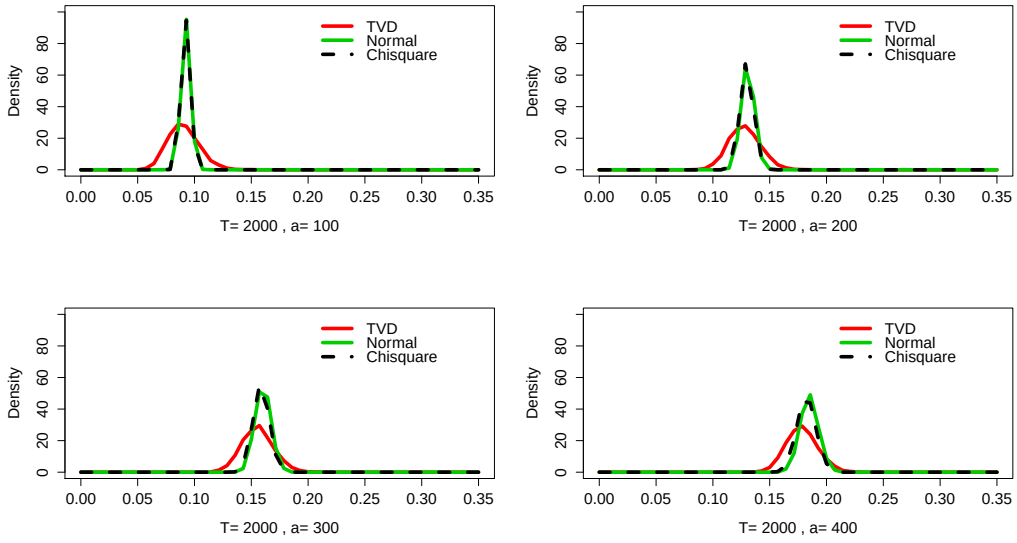
So, under H_0 and using that $\displaystyle\int f(\omega)\mathrm{d}(\omega) = \sigma^2$,

$$\begin{aligned} 2\mathbb{E}(\tilde{d}_{TV}) &= \sigma^2\Sigma\sqrt{\frac{2a}{\pi T}} = K_1\sqrt{\frac{a}{T}}, \\ 4\mathrm{Var}(\tilde{d}_{TV}) &= \int f^2(\omega)\mathrm{d}(\omega)\Sigma^2\left(1 - \frac{2}{\pi}\right)\frac{a}{T} = K_2\frac{a}{T}, \end{aligned} \tag{2.32}$$

where $\tilde{d}_{TV}$ denotes the normal approximation of $\hat{d}_{TV}$.
So, the factor a/T together with the first two spectral moments determine the dispersion and the mean of the normal approximation. It is true that both objects go to zero when a/T goes to zero. However, for T fixed, when

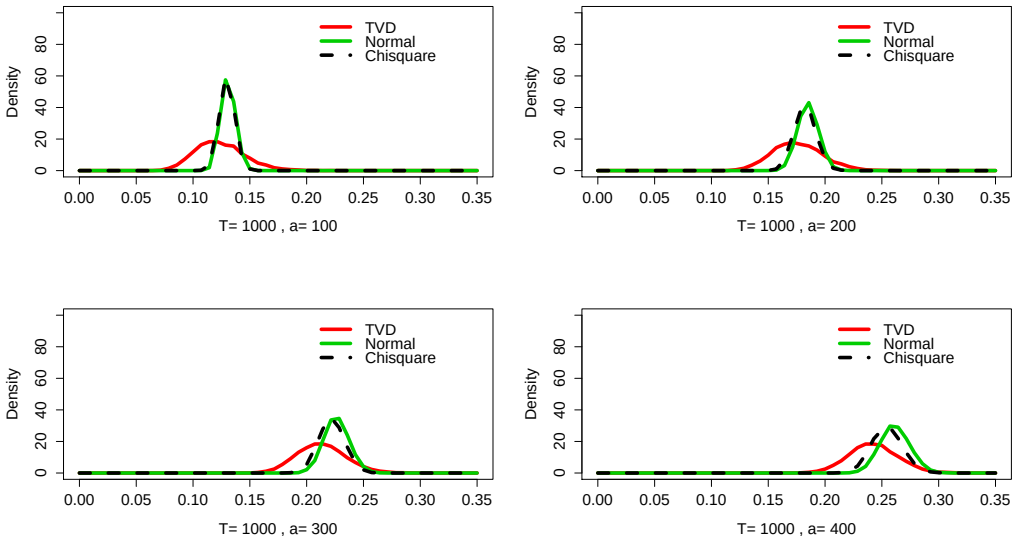


(a) $T = 1000$

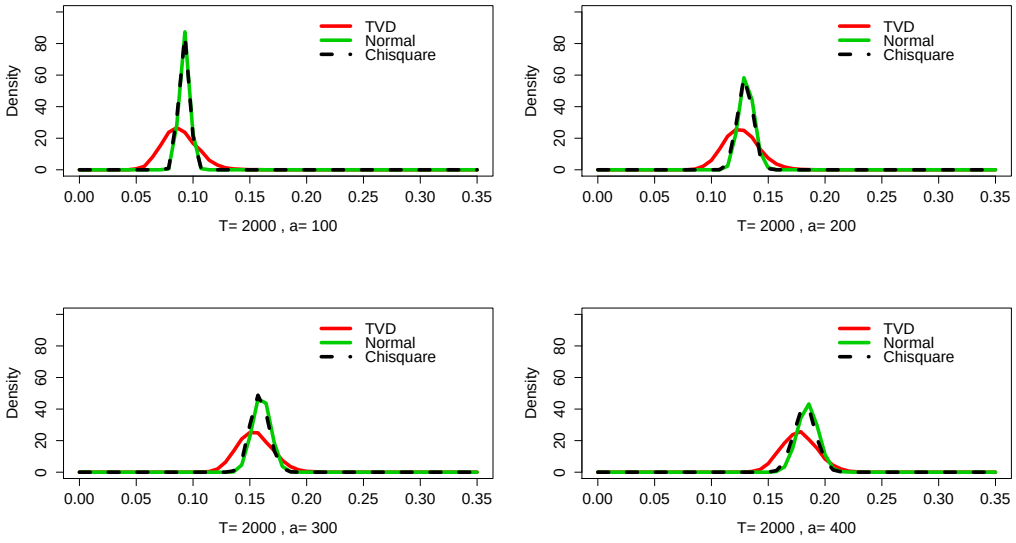


(b) $T = 2000$

Figure 2.9: Simulation results for AR(1) with small sample size and different values of the bandwidth, $T = 1000, 2000$, $a = 100, 200, 300, 400$.



(a) $T = 1000$



(b) $T = 2000$

Figure 2.10: Simulation results for AR(2) with small sample size and different values of the bandwidth,, $T = 1000, 2000$, $a = 100, 200, 300, 400$.

a is closer to T we increase the dispersion but also the mean, and a bias appears. This is the phenomenon we observe in the simulations.

We would like to increase the dispersion but not the mean, and also we would like to know how much we should increase it. To prove the convergence of the smoothed periodogram the following inequality is used (Brockwell and Davis, 2006),

$$\left(\sum \boldsymbol{\beta}_T^2(\omega_j)\right)^{-1} \mathrm{Var}(\hat{f}(\omega)) \leq f^2(\omega) + \left(\sum \boldsymbol{\beta}_T^2(\omega_j)\right)^{-1} o\left(\sum \boldsymbol{\beta}_T^2(\omega_j)\right) + c_2\left(\tfrac{2a+1}{T}\right),$$

where c_2 is a constant and $\displaystyle\sum_{|j|<a} \boldsymbol{\beta}_T^2(\omega_j) \approx \frac{a}{T}\int_{-1}^{1} \beta^2(u)\mathrm{d}u$. This inequality and (2.32) motivate the following proposal.

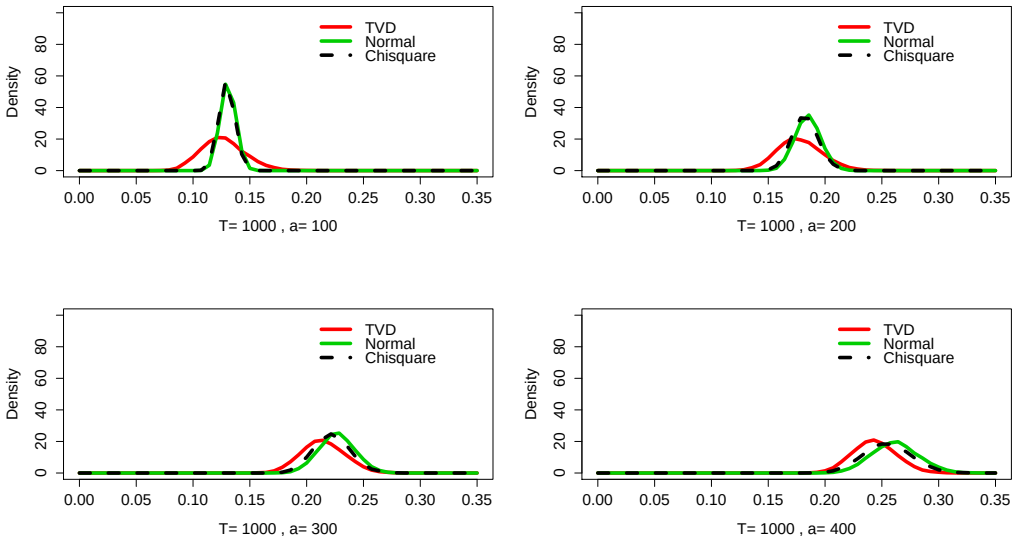
Consider a transformation of the random variable $\tilde{d}_{TV}$, that approximates $\hat{d}_{TV}$ by the following function,

$$\left(1 + \frac{2a+1}{T}\right)\tilde{d}_{TV} - \left(\frac{2a+1}{T}\right)\mathbb{E}(\tilde{d}_{TV}). \tag{2.33}$$

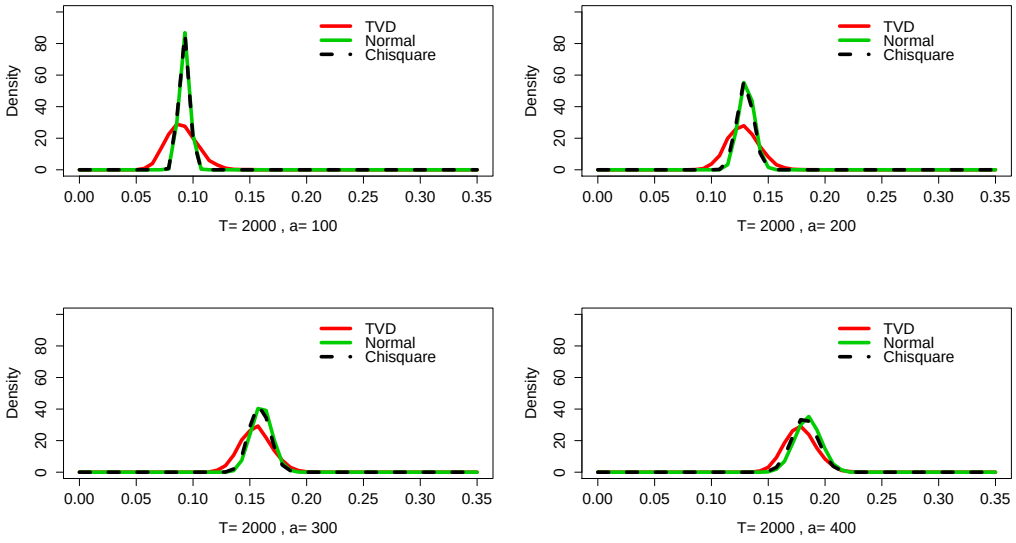
The results obtained are presented in Figures 2.11 and 2.12. The transformed approximations, for values of a bigger than 300, capture the right dispersion of the distribution and reduce the bias. However, the approximations are not completely accurate. This is to be expected since we have "small" values of T.

Bootstrapping. Now, we explore the approximation using the bootstrap procedure. The simulation setting in this case is $T = 1000, 2000$ and $a = 100, 150, 200, 250$. We consider the AR(1) and AR(2) processes as before. In this case we take one pair of samples from $[X_1(t), X_2(t)]$, then we draw a bootstrap sample based on them. Finally, we compare this bootstrap density with a density obtained by $\hat{d}_{TV}$ from different replicates of $[X_1(t), X_2(t)]$. Figures 2.13 and 2.14 show the results.

The bootstrap density is a good approximation to the $\hat{d}_{TV}$ density. It doest not depend on a, in the sense that under any value of a the approximations are very close to the empirical. The performance of the bootstrap is equally precise for both processes.

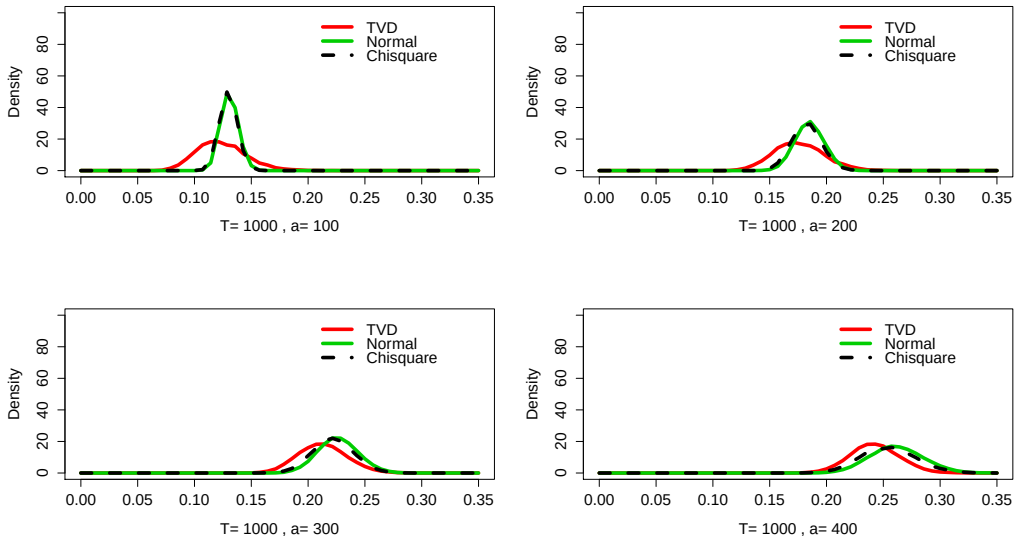


(a) $T = 1000$

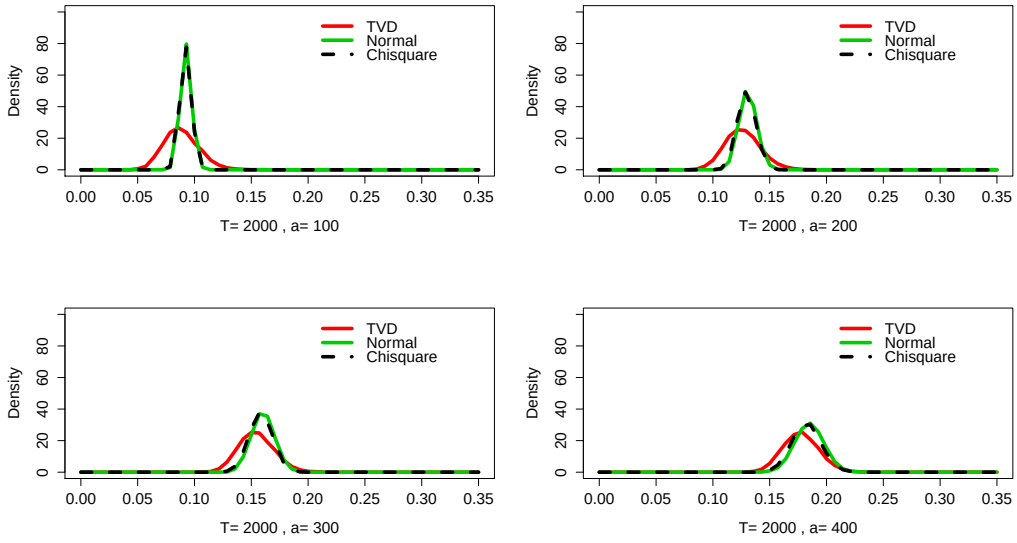


(b) $T = 2000$

Figure 2.11: Results using the transformed values for AR(1), $T = 1000, 2000$, $a = 100, 200, 300, 400$.

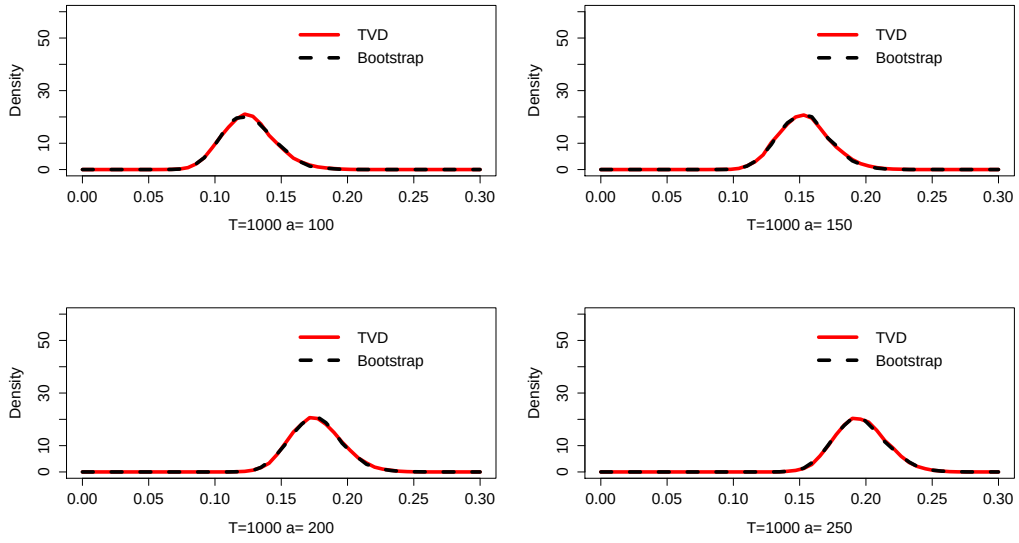


(a) $T = 1000$

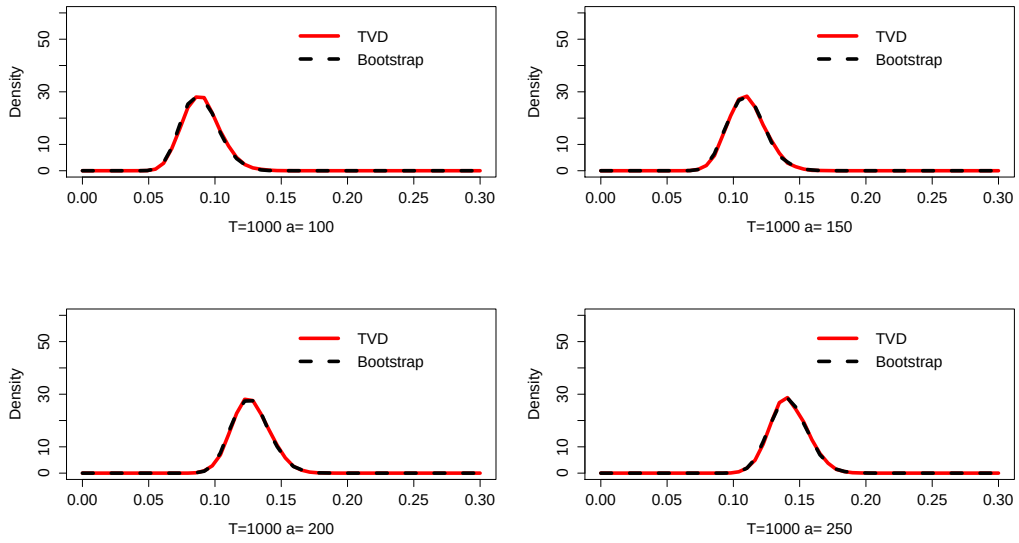


(b) $T = 2000$

Figure 2.12: Results using the transformed values for AR(2), $T = 1000, 2000$, $a = 100, 200, 300, 400$.

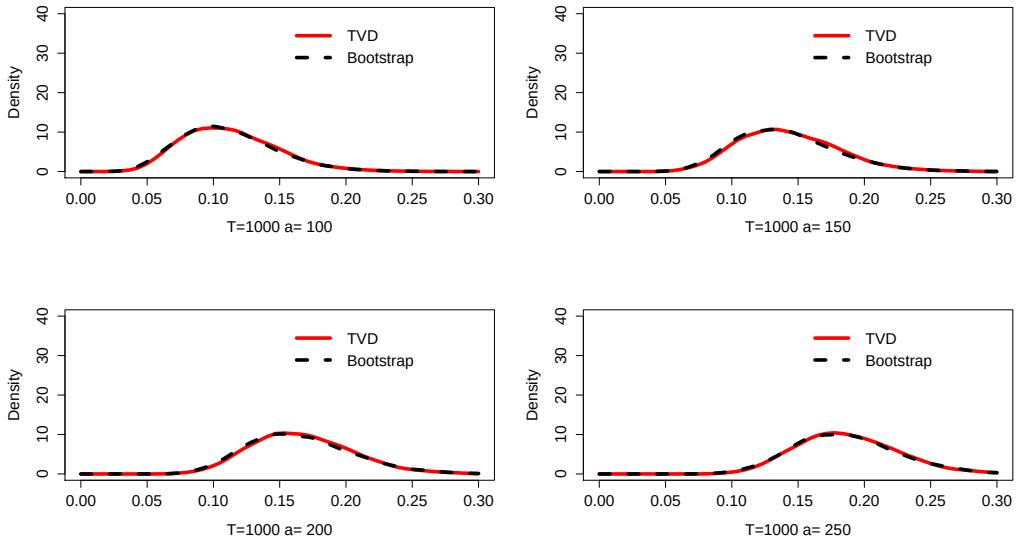


(a) $T = 1000$

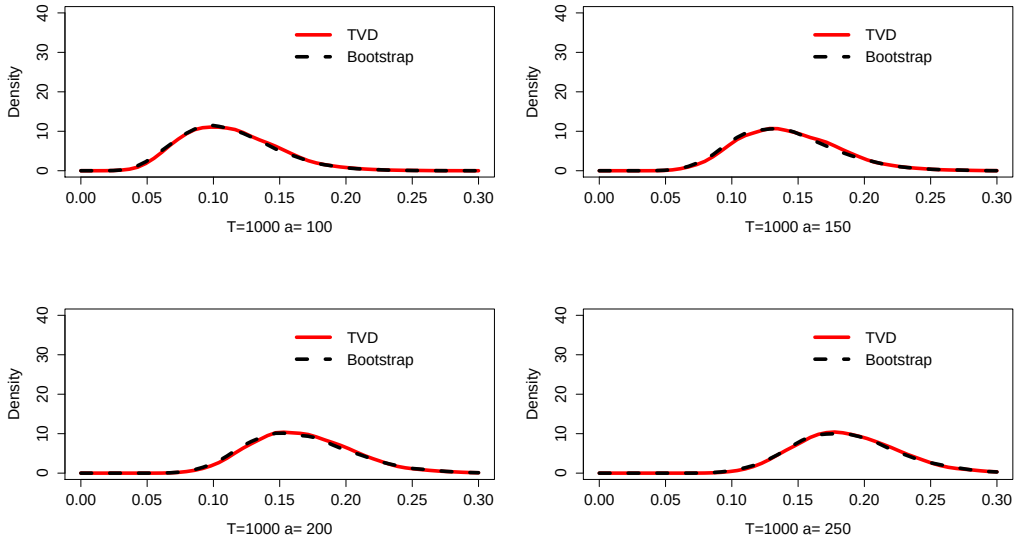


(b) $T = 2000$

Figure 2.13: Results using bootstrap for AR(1), $T = 1000, 2000$, $a = 100, 150, 200, 250$.



(a) $T = 1000$



(b) $T = 2000$

Figure 2.14: Results using bootstrap for AR(2), $T = 1000, 2000$, $a = 100, 150, 200, 250$.

2.4.2 Significance level and power of the test

Given two time series, $X_1(t)$ and $X_2(t)$, we establish the following hypothesis test.
Hypothesis:

$$H_0 : f^{X_1}(\omega) = f^{X_2}(\omega),\ \forall\omega \quad vs \quad H_A : \exists\omega \text{ such that } f^{X_1}(\omega) \neq f^{X_2}(\omega).$$

Test statistic:

$$\hat{d}_{TV} = \frac{1}{2T}\sum_{k=1}^{T}\left|\hat{f}_N^{X_1}\left(\frac{k}{T}-\frac{1}{2}\right) - \hat{f}_N^{X_2}\left(\frac{k}{T}-\frac{1}{2}\right)\right|.$$

In both cases, to determine the rejection region or to compute the p-value, we need an approximation of the distribution of our test statistic $\hat{d}_{TV}$. We use one of the following distributions: 1) Normal approximation, 2) Transformed Normal approximation, 3) Chi-square approximation, 4) Transformed Chi-square approximation and 5) Bootstrapping.

We will approximate the distribution of our test statistic, so the first thing we would like to verify is the significance level. In other words, with a fixed value c_α^i we want to verify how close is $\mathbb{P}(\hat{d}_{TV} > c_\alpha^i)$ to α, where $i = 1, \ldots, 5$ indicates which approximation we are using. Also, we want to study the power of the test.

Significance Level. To explore the first property we shall consider the AR(2) process with peak frequency at .3 Hz, $(\phi_1, \phi_2) = (-.47, -.6)$. We use $T = 1000, 2000, 5000$, $a = 100, 200, 300, 400$, and 1000 replicates to approximate the distributions and 1000 replicates to study the test performance. The pair (T, a) is specified in Table 2.3.

The simulation procedure is the following.

- We draw two time series from the AR(2) process with the same parameters.
- We estimate $\hat{f}_N^{X_1}$, $\hat{f}_N^{X_2}$ and the "true" spectra as $\hat{f}_N = \dfrac{\hat{f}_N^{X_1} + \hat{f}_N^{X_2}}{2}$.
- We estimate the quantiles c_α^i using, $i = 1$ - Normal approximation, $i = 2$ - Transformed Normal approximation, $i = 3$ - Chi-square approximation, $i = 4$ - Transformed Chi-square approximation and $i = 5$ - Bootstrapping.

Under H_0

	α	Normal	Transformed Normal	Chi-square	Transformed Chi-square	Bootstrapping
$T = 1000$	0.01	0.134	0.117	0.143	0.117	0.012
$a = 100$	0.05	0.185	0.164	0.189	0.169	0.050
	0.1	0.213	0.189	0.223	0.202	0.100
$T = 1000$	0.01	0.075	0.043	0.086	0.048	0.012
$a = 200$	0.05	0.126	0.074	0.147	0.086	0.05
	0.1	0.158	0.11	0.178	0.135	0.097
$T = 2000$	0.01	0.128	0.101	0.133	0.107	0.009
$a = 200$	0.05	0.174	0.149	0.187	0.159	0.042
	0.1	0.209	0.183	0.226	0.201	0.077
$T = 2000$	0.01	0.119	0.072	0.133	0.085	0.01
$a = 300$	0.05	0.166	0.133	0.18	0.147	0.056
	0.1	0.191	0.164	0.206	0.183	0.106
$T = 5000$	0.01	0.202	0.193	0.215	0.200	0.014
$a = 300$	0.05	0.253	0.240	0.262	0.252	0.053
	0.1	0.273	0.265	0.292	0.277	0.096
$T = 5000$	0.01	0.186	0.161	0.199	0.175	0.010
$a = 400$	0.05	0.232	0.214	0.252	0.230	0.056
	0.1	0.260	0.245	0.279	0.262	0.104
$T = 5000$	0.01	0.167	0.129	0.188	0.159	0.007
$a = 500$	0.05	0.218	0.191	0.234	0.210	0.050
	0.1	0.243	0.226	0.268	0.248	0.102

Table 2.3: Proportion of rejection H_0 using different approximated distribution of $\hat{d}_{TV}$

- Finally we compute $\hat{d}_{TV}$ and compare with c_α^i.

Table 2.3 shows the proportion of times that the null hypothesis is rejected using the critical value associated to each approximation. If we observe the theoretical approximation, the transformed values, compared to the non transformed, have proportions closer to α. As we expect, for the theoretical approximation the value of a has an influence on the proportion of rejection, bigger values of T need bigger values of a. On the other hand, the bootstrap procedure outperforms the rest in all cases. The proportion of rejection using the bootstrap procedure is almost equal to α, and is not influenced by the choice of a.

Power. Now, we draw a time series from the AR(2) process but with different parameters. Figure 2.15 shows the spectrum for X_1, the continuous black curve, and for X_2 we use three different cases, the dotted curves. So, we fix the spectra for X_1 and we use one of the others for X_2. They are very close and we would like to see how many false positive there will be as the spectra get closer.

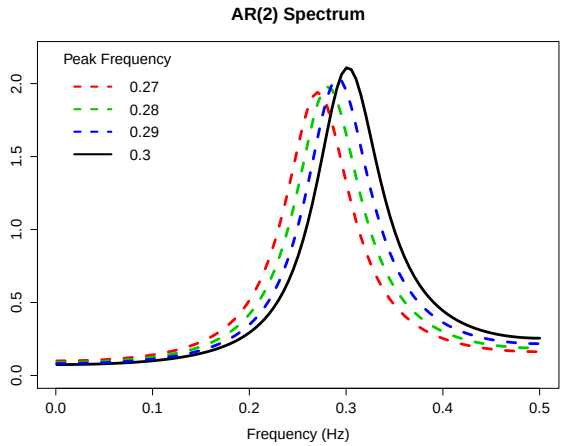


Figure 2.15: Spectra under H_A.

We use $T = 1000, a = 200$ (Table 2.4) and $T = 2000, a = 300$ (Table 2.5), and 1000 replicates to approximate the distributions and 1000 replicates to study the test performance.

In all cases, the power decreases when the spectra are closer. The Chi-squared approximation has the biggest power. If we compare the power of the theoretical and the transformed approximations, the transformation does not improve the power. This fact is a consequence of the subestimation of the dispersion of the distribution of $\hat{d}_{TV}$.
The power in the bootstrap case is closer to the theoretical approximation when the spectrum of X_2 has the peak frequency at .27. When the spectra are closer the power decreases but it is still acceptable (around .7) when the peak frequency is at .28.

The comparison between the power of the bootstrap procedure and the asymptotic approximations is not completely fair, since the significance level when one uses the asymptotic distributions is bigger than the fixed level. Since the bootstrap procedure is the only option that preserves the significance level, it will be the best option to use in practice even when the power could be low.

Under H_A $T = 1000$ $a = 200$

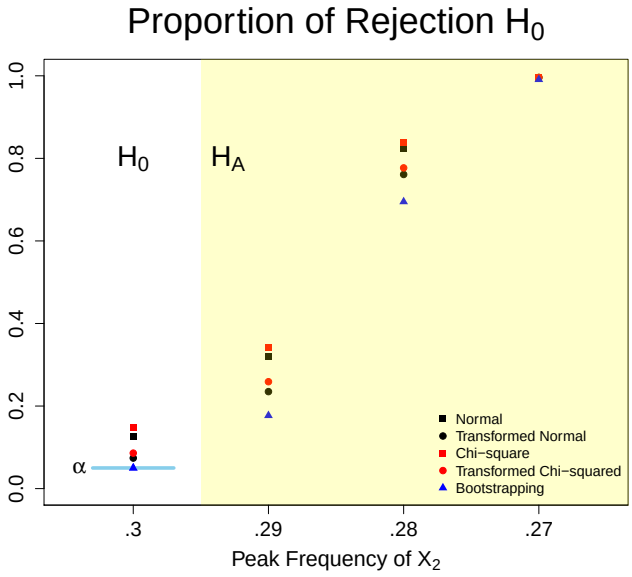
Peak Frequency of X_2 spectra	α	Normal	Transformed Normal	Chi-square	Transformed Chi-square	Bootstrapping
0.27	0.01	0.996	0.988	0.996	0.99	0.965
0.27	0.05	0.997	0.996	0.997	0.996	0.991
0.27	0.1	0.997	0.997	0.998	0.997	0.997
0.28	0.01	0.758	0.658	0.774	0.682	0.467
0.28	0.05	0.825	0.761	0.839	0.777	0.695
0.28	0.1	0.847	0.81	0.875	0.832	0.792
0.29	0.01	0.236	0.146	0.257	0.159	0.068
0.29	0.05	0.319	0.235	0.342	0.259	0.177
0.29	0.1	0.367	0.305	0.416	0.326	0.276

Table 2.4: Proportion of rejection H_0 under H_A. $T = 1000$ and $a = 200$.

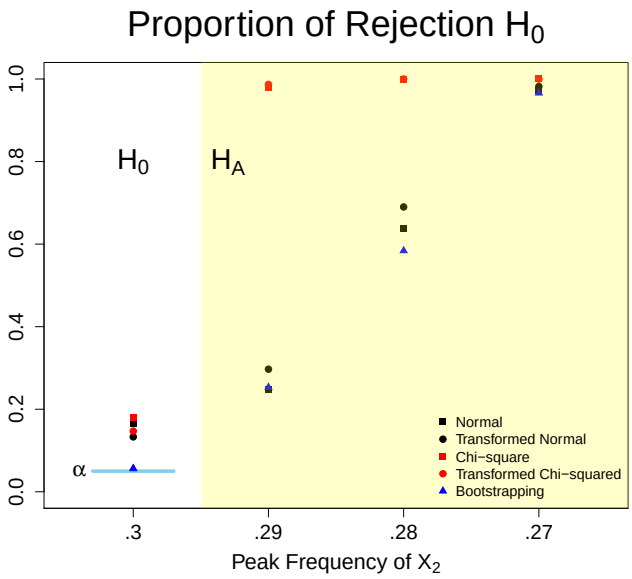
Under H_A $T = 2000$ $a = 300$

Peak Frequency of X_2 spectra	α	Normal	Transformed Normal	Chi-square	Transformed Chi-square	Bootstrapping
0.27	0.01	0.951	0.964	1.000	1.000	0.902
0.27	0.05	0.973	0.982	1.000	1.000	0.966
0.27	0.1	0.983	0.986	1.000	1.000	0.984
0.28	0.01	0.517	0.576	0.991	0.993	0.332
0.28	0.05	0.637	0.690	0.999	1.000	0.584
0.28	0.1	0.701	0.746	1.000	1.000	0.709
0.29	0.01	0.112	0.150	0.898	0.915	0.037
0.29	0.05	0.248	0.297	0.980	0.987	0.253
0.29	0.1	0.199	0.242	0.966	0.971	0.161

Table 2.5: Proportion of rejection H_0 under H_A. $T = 2000$ and $a = 300$



(a) $T = 1000$, $a = 200$



(b) $T = 2000$, $a = 300$

Figure 2.16: Proportion of rejection of H_0 considering a level $\alpha = 0.05$ and different approximations of the distribution of $\hat{d}_{TV}$. The x-axis represents the peak frequency (T_p) of the spectra used to draw $X_2(t)$, when $T_p = .3$, $X_1(t)$ and $X_2(t)$ have the same spectral density.

2.5 Discussion

In comparison with other similarity measures, the total variation distance has some desirable properties. The intuition and easy interpretation is one of them. Also, contamination models give an interpretation of the distance as the level of similarity. It is important to note that we use the total variation distance to compare continuous functions, since the total variation distance is not useful to compare discrete with continuous functions.

We explored the statistical properties of the estimator of the total variation distance, $\hat{d}_{TV}$. Two approximations of the distribution are proposed, using Gaussian or Chi-squared variables, and a transformation of them is introduced in the case of small samples. In the simulation study, the test based on these distribution has a bigger significance level than the nominal α. The transformations gave a closer significance level to α, however, they are not sufficiently precise for "small" T.

As an alternative, we propose a bootstrap algorithm and the results are very good. The bootstrap outperforms the asymptotic methods and the significance level is almost equal to α. It has the limitation of low power when the spectral densities are very close. In general, the bootstrap procedure is the best option to approximate the distribution of $\hat{d}_{TV}$, under the null hypothesis.

The developed theory can be extended to the multivariate case. Another posible extension is to consider the distances between some operator of the spectral density such that the first or second derivative of the spectra, this would be useful in some applications.

Chapter 3

Clustering Methods

Our main goal is to detect changes in spectra and the previous chapter explores the proposal of considering the TV distance as a similarity measure between spectra. As was mentioned in the introduction, several methods for detecting instantaneous breaks in time series have been proposed, but they do not produce good results when the changes are slow. In this situation it is convenient to change the point of view from detecting change-points to determining time intervals during which the spectra are similar, in the sense that their TV distance is small. If one considers that time series that have similar spectral densities also share similar properties, one could think about them as a group. Taking this into account, clustering methods are a natural approach. Clustering based on spectral densities will be intuitive in many applications.

In general, clustering is a procedure whereby a set of unlabeled data is divided into groups so that members of the same group are similar, while members of different groups differ as much as possible. Our goal is to develop a method that produces groups or clusters consisting of time series having similar spectral representation.

The subject of time series clustering is an active research area with applications in many fields. Frequently, finding similarity between time series plays a central role in the applications. In fact, time series clustering problems arise in a natural way in a wide variety of fields, including economics, finance, medicine, ecology, environmental studies, engineering, and many others. This is not an easy task, since it requires a notion of similarity between time series. Liao (2005) and Caiado et al. (2015) give a revision of the field and Montero and Vilar (2014) present an R

package (TSclust) for time series clustering with a wide variety of alternative procedures. According to Liao (2005), there are three approaches to time series clustering: methods based on the comparison of raw data, feature-based methods, where the similarity between time series is gauged through features extracted from the data, and methods based on parameters from models adjusted to the data.

The first approach, comparison of raw data, will be very complicated when we have long time series, since it becomes a computational problem. The third approach, based on parameters, is one of the most frequently used, however, it has the limitation of considering a specific parametric model.

Our proposals are feature-based and the spectral density of the time series is considered the central feature for classification purposes. The resulting clusters will be similar in the sense that the time series in a cluster will have similar spectral density. This will have an interpretation depending on the application, Chapter 4 presents two different cases and the interpretation for each one.

To build a clustering method the first question is how to measure the similarity between spectral densities. We propose the use of the total variation distance as a measure of similarity. Then, we need a clustering algorithm, and we use a hierarchical algorithm with classical linkage functions as our first proposal.

However, hierarchical clustering algorithms with linkage functions (such as complete, average, Ward, and so on) are based on geometrical ideas where the distances between new clusters and old ones are computed by a linear combination of the distance of their members, which may not be meaningful for clustering time series since these linear combinations may not have a meaning in terms of the spectral densities. So, our second proposal considers a new clustering algorithm, which takes advantage of the spectral theory. We propose the Hierarchical Spectral Merger algorithm, which is a modification of the classical hierarchical algorithms. The main difference is the consideration of a new representative, i.e. a new estimation of the spectral density for an updated cluster. This is intuitive and the updated spectral estimates are smoother, less noisy and hence give better estimates of the TV distance.

We explain each proposal in detail and compare through simulation studies their performance.

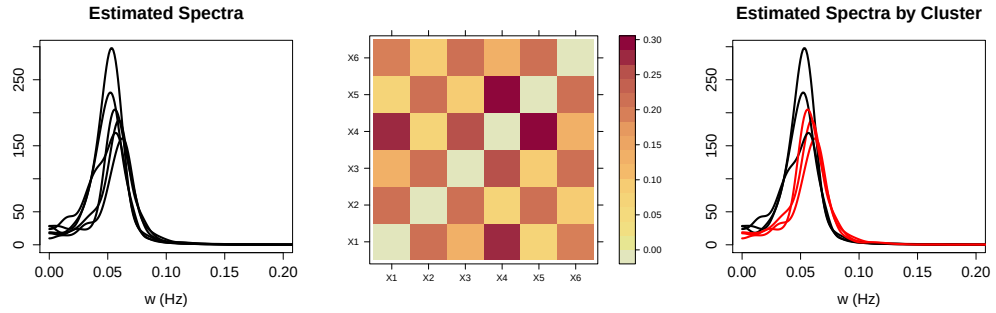
3.1 TV distance in a clustering method

There are two general families of clustering algorithms: partitioning and hierarchical. Among partitioning algorithms, K-means and K-medoids are the two more representative, and for the hierarchical clustering algorithms, the main examples are agglomerative with single-linkage or complete-linkage (Xu and Wunsch, 2005).

We consider the hierarchical algorithm because it accepts as input the distances between objects. The algorithm starts with a dissimilarity matrix and builds each cluster giving preference to the closest candidates. Distances between new and old clusters are required during the algorithm and they are calculated using the link functions. In complete link clustering, the distance is equal to the longest distance from any member of one cluster to any member of the other cluster. In average link clustering, the distance is equal to the average distance from any member of one cluster to any member of the other cluster.

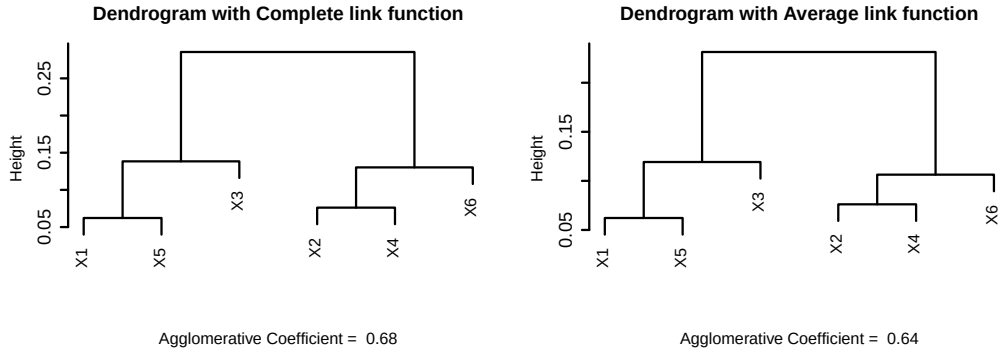
Let $X_i = (X_i(1), \ldots, X_i(T))$ be a set of time series, $i = 1, \ldots, N$. The first proposed procedure will be as follows.
Step 1. Estimate the spectral density for each time series using the smoothed periodogram.
Step 2. To calculate the dissimilarity matrix, the TV distance between the normalized spectra is used.
Step 3. The dissimilarity matrix is used for the hierarchical algorithm, with the complete and average link functions.
Step 4. As a result, a clustering dendrogram is obtained for the data set in which the distance between groups is represented by the length of the segments, and one can decide to cut the dendrogram according to a fixed value of the distance or to the number of clusters desired.

Example 3.1. Consider two different AR(2) models with spectra concentrated at 0.05 Hz and 0.06 Hz, respectively. We simulate three time series for each process, each one consisting of 1000 points with a sampling frequency of 1 Hz, being X_1, X_3, X_5 from the first process and X_2, X_4, X_6 from the second process. Figure 3.1(a) shows the estimated spectra for each series. We compute the dissimilarity matrix with the TV distance and the values are shown in 3.1(b). These values are not big since the spectra are close. When we apply the hierarchical algorithm with complete and average link



(a) (b) (c)

Figure 3.1: (a) Estimated spectra for Example 3.1, (b) Dissimilarity Matrix, computed usig the TV distance, and (c) Clustering result using either the complete or average link functions.



(a) (b)

Figure 3.2: Dendrograms obtained for Example 3.1 using (a) the complete link function and (b) the average link function.

functions, we obtain the dendrogram plots in Figures 3.2(a) and (b). If we consider two groups, Figure 3.1(c) shows in different color (red and black) which one belongs to each cluster. The method is able to recover the original groups. Note that new distances obtained using the complete link are bigger than those obtained with the average link. It could be possible that for detecting small changes, the complete link could have better results.

In the simulation study, we will study the case when the number of groups, k, is unknown.

3.2 Hierarchical spectral merger (HSM) method

As an alternative to the previous proposal, a new time series clustering algorithm was developed. The Hierarchical Spectral Merger (HSM) method uses the TV distance as a dissimilarity measure as before but new clustering procedures are introduced. The algorithms proposed are a modification of the usual agglomerative hierarchical procedure, taking advantage of the spectral point of view for the analysis of time series.

The hierarchical spectral merger algorithm has two versions: the first, known as *single version*, updates the spectral estimate of the cluster from a concatenation of the time series; and the second, known as *average version*, updates the spectral estimate of the cluster from a weighted average of the spectral estimate obtained from each signal in the cluster.

Hierarchical Spectral Merger Algorithm. Let $X_i = (X_i(1), \ldots, X_i(T))$ be a set of time series, $i = 1, \ldots, N$. The procedure starts with N clusters, each cluster being a single signal.
Step 1. Estimate the spectral density for each cluster using the smoothed periodogram, then each cluster will be represent by a common spectral density $f_j(\omega)$, $j = 1, \ldots, k$ (number of clusters).
Step 2. Compute the TV distance between their spectra.
Step 3. Find the two clusters that have lowest TV distance, save this value as a characteristic.
Step 4. Merge the signals in the two closest clusters and replace the two clusters by this new one.
Step 5. Repeat Steps 1-4 until there is only one cluster left.

The characteristic saved in Step 3 represents the "cost" of joining two clusters, i.e., having $k - 1$ clusters vs k clusters. If a significantly large

Algorithm:

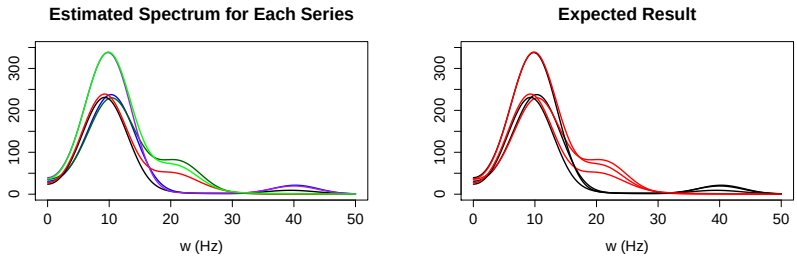
1. Initial clusters: $\mathbf{C} = \{C_i\}$, $C_i = X_i$, $i = 1, \ldots, N$
   Dissimilarity matrix entry between clusters i and j,
   $$D_{ij} = d(C_i, C_j) := d_{TV}(\widehat{f}_i, \widehat{f}_j),$$
   $\widehat{f}_i$ is the estimated spectra using the signals in each cluster.
2. **for** k *in* $1 : N-1$
3. $(i_k, j_k) = \underset{ij}{\operatorname{argmin}} D_{ij}$; $min_k = \min_{ij} D_{ij}$ #Find the closest clusters
4. $C_{new} = C_{i_k} \cup C_{j_k}$ #Join the closest clusters
5. $D^{new} = D \setminus \{D_{i_k \cdot} \cup D_{j_k \cdot} \cup D_{\cdot i_k} \cup D_{\cdot j_k}\}$ #Delete rows and columns i_k, j_k
6. **for** j *in* $1 : N-k-1$
7. $D^{new}_{(N-k)j} = D^{new}_{j(N-k)} = d_{TV}(C_{new}, C_j)$ #Compute new distances
8. **end**
9. $D = D^{new}$; $\mathbf{C} = (\mathbf{C} \setminus \{C_{i_k}, C_{j_k}\}) \cup C_{new}$ #New matrix D and new clusters
10. **end**

Table 3.1: Hierarchical Merger Algorithm proposed using the total variation distance and the estimated spectra.

value is observed, then it may be reasonable to choose k clusters instead of $k-1$. When two clusters merge, there are two options, either (1) for the single version, we concatenate both signals and compute the smoothed periodogram with the concatenated signal; or (2) for the average version, we take the weighted average over all the estimated spectra for each signal in the cluster as the new estimated spectra.

Both algorithms compute the TV distance between the new cluster and the old clusters based on updated estimated spectra, which is the main difference with classical hierarchical algorithms. While a hierarchical algorithm has a dissimilarity matrix of size $N \times N$ during the whole algorithm, the proposed method reduces this size to $(N-k) \times (N-k)$ at the k-th iteration. Table 3.1 gives a summary of the hierarchical merger algorithm.

Example 3.2. To illustrate the HSM method, consider two different AR(2) models with their spectra concentrated at 10 Hz, however, one also contains



(a) (b)

Figure 3.3: Estimated spectra. (a.) Different colors correspond to different time series, (b.) Red spectra are from the AR(2) model with activity at alpha (8-12 Hz) and beta (12-30 Hz) bands and black spectra are from the AR(2) model with activity at alpha and gamma (30-50 Hz) bands.

power at 21 Hz while the other has power at 40 Hz. We simulate three time series for each process, 10 seconds of each one with a sampling frequency of 100 Hz ($t = 1, \ldots, 1000$). Figure 3.3(a) shows the estimated spectra for each series and Figure 3.3(b) shows by different colors (red and black) which one belongs to the first or second process. If we only look at the spectra, it is hard to recognize the number of clusters and their memberships. We probably could not identify some cases, like the red and purple spectra.

The dynamics of the HSM method is shown in Figure 3.4. We start with six clusters; at the first iteration we find the closest spectra, represented in Figure 3.4(a) with the same color (red). After the first iteration we merge these time series and get 5 estimated spectra, one per cluster, Figure 3.4(b) shows the estimated spectra where the new cluster is represented by the dashed red curve. We can follow the procedure in Figures 3.4(c), (d), (e) and (f). In the end, the proposed clustering algorithm reaches the correct solution, Figures 3.4(g) and 3.3(b) coincide. Also, the estimated spectra for the two clusters, Figure 3.4(h), is better than any of the initial spectra and we can identify the dominant frequency bands for each cluster.

We developed the *HSMClust* package written in R that implements our proposed clustering method. The package can be downloaded from `http://ucispacetime.wix.com/spacetime#!project-a/cxl2`.
The principal function is called *HSM*, which executes the HSM method given a matrix X, which has the signals by column. *HSMClust* has also some other useful functions. One of them is the *Sim.Ar* function, which draws

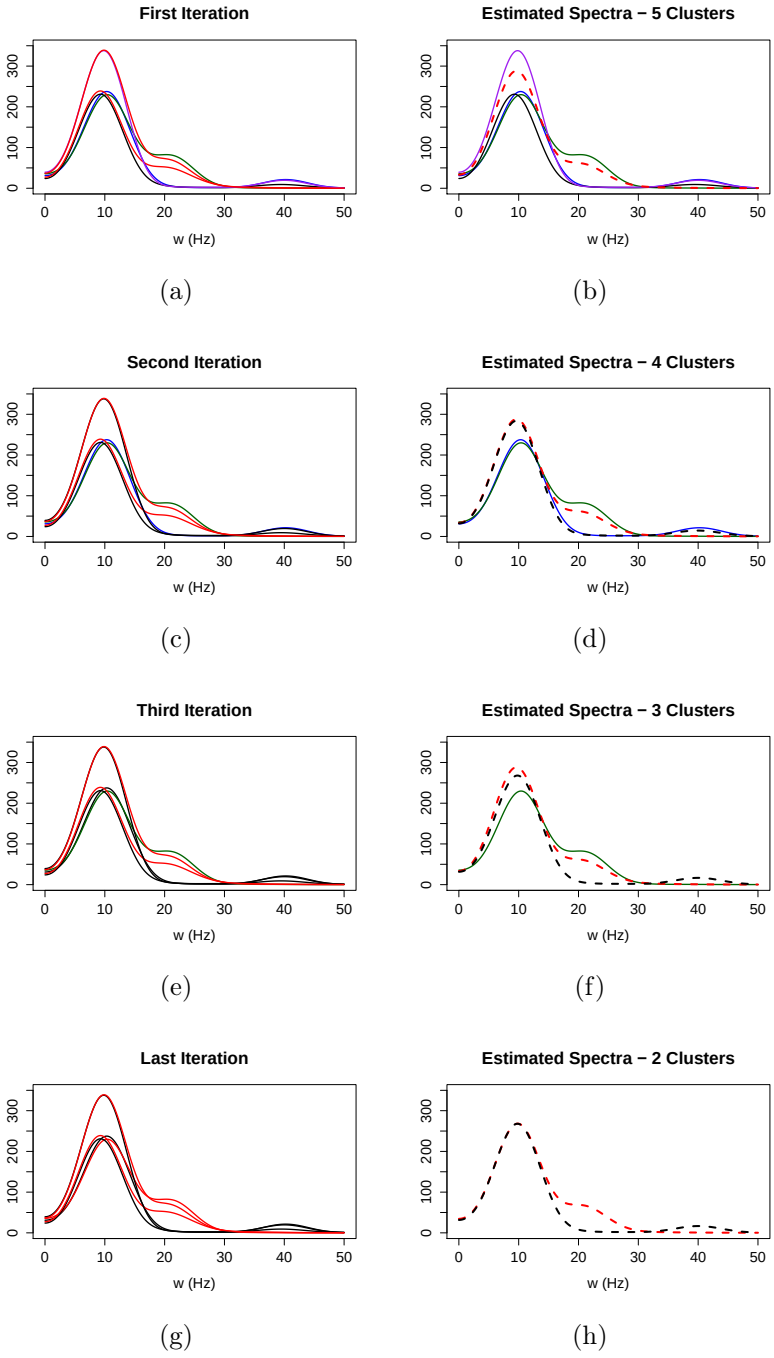


Figure 3.4: Dynamic of the hierarchical merger algorithm. (a), (c), (e) and (g) show the clustering process for the spectra. (b), (d), (f) and (h) show the evolution of the estimated spectra, which improves when we merge the series in the same cluster.

an AR(2) process reparametrized as a function of the norm of the root of its characteristic polynomial (M) and the peak frequency of the spectrum (η). These parameters determine the shape of the spectral density (peak and sparseness).

3.3 TV distance and other dissimilarity measures

In this section we study the performance of the proposed methods and compare them with methods based on other distances.
First, we explain the simulation methods based on the spectrum that we will use in the experiments. Then, we present the results of the experiments, assuming that the number of clusters is known. Finally, we explore the case of unknown number of clusters and possible criteria to choose the number of clusters.

This method can be applied using the *agnes* function of the R package *cluster*.

3.3.1 Simulation of a process based on the spectral density

There exist several methods to simulate a time series based on the spectral density, usually they depend on a model for the spectral density. We consider two of them, based on a parametric spectral density and based on AR(2) processes.

Simulation based on a parametric family of spectral densities. There exist several parametric families of spectra of frequent use in Oceanography and they have an interpretation in terms of the behavior of sea waves (Ochi, 1998). Motivated by the applications we will present in the next chapter, we will simulate time series (Gaussian process) using spectra from several of these families. This methodology is already implemented by Brodtkorb et al. (2011) in the WAFO toolbox for MATLAB.

WAFO has a routine for simulation of (transformed) Gaussian processes and their derivatives, using a technique of circulant embedding of the covariance matrix proposed by Dietrich and Newsam (1997). More

traditional spectral simulation methods (FFT) are also implemented, see the WAFO tutorial for more details.

An example of a group of parametric densities is the JONSWAP (Joint North-Sea Wave Project) spectral family. This is a parametric family of spectral densities which is frequently used in Oceanography, and is given by the formula

$$S(\omega) = \frac{g^2}{\omega^5} \exp(-5\omega_p^4/4\omega^4)\gamma^{\exp(-(\omega-\omega_p)^2/2\omega_p^2 s^2)}$$

where g is the acceleration of gravity, $s = 0.07$ if $\omega \leq \omega_p$ and $s = 0.09$ otherwise; $\omega_p = \pi/T_p$ and $\gamma = \exp(3.484(1 - 0.1975(0.036 - 0.0056T_p/\sqrt{H_s})T_p^4/(H_s^2)))$. The parameters for the model are the significant wave height H_s, which is defined as 4 times the standard deviation of the time series, and the spectral peak period T_p, which is the period corresponding to the modal frequency of the spectrum. This spectral family was empirically developed after analysis of data collected during the Joint North Sea Wave Observation Project, JONSWAP, (Hasselmann et al., 1973). It is a reasonable model for wind-generated seas when $3.6\sqrt{H_s} \leq T_p \leq 5\sqrt{H_s}$.

Simulation based on AR(2) processes. Consider the second order auto-regressive model which is defined as

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \epsilon_t, \tag{3.1}$$

where ϵ_t is a white noise process. The characteristic polynomial for this model is $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$. The roots of the polynomial equation indicate the properties of the oscillations. If the roots, denoted z_0^1 and z_0^2 are complex-valued then they have to be complex-conjugates, i.e., $z_0^1 = \overline{z_0^2}$. These roots have a polar representation

$$|z_0^1| = |z_0^2| = M, \qquad \arg(z_0) = \frac{2\pi\eta}{F_s}, \tag{3.2}$$

where F_s denotes the sampling frequency; M is the amplitude or magnitude of the root ($M > 1$ for causality); and η is the frequency index. The spectrum of the AR(2) process with polynomial roots above will have peak frequency at η. The peak becomes broader as $M \to \infty$, and it becomes narrower as

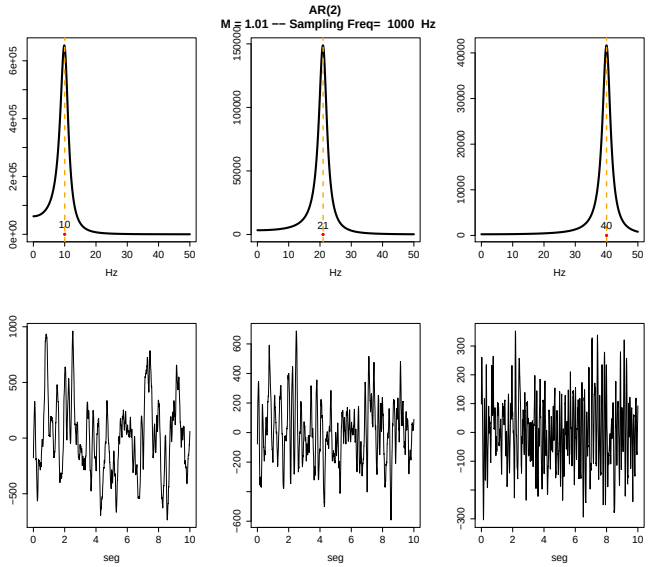


Figure 3.5: Top: Spectra for the AR(2) process with different peak frequency; $\eta = 10, 21, 40$. Bottom: Realization from the corresponding AR(2) process.

$M \to 1^+$.
Then, given (η, M, F_s) we take

$$\phi_1 = \frac{2\cos(w_0)}{M} \qquad \text{and} \qquad \phi_2 = \frac{-1}{M^2}, \tag{3.3}$$

where $w_0 = \frac{2\pi\eta}{F_s}$. If one computes the roots of the characteristic polynomial with the coefficients in (3.3), they satisfy (3.2). To illustrate the type of oscillatory patterns that can be observed in time series from processes with corresponding spectra, we plot in Figure 3.5 the spectra (top) for different values of η, $M = 1.1$ and $F_s = 100Hz$; and the generated time series (bottom). Larger values of η gives rise to faster oscillation of the signal.

3.3.2 Comparative study

Pértega Díaz and Vilar (2010) proposed two simulation tests to compare the performance of several dissimilarity measures for time series clustering. Our goal in this section is to compare the performance of the TV distance with those that were based on the spectral density and had good results in Pértega and Vilar's experiments. In addition, we use the distance based

on the cepstral coefficients (the Fourier coefficients of the expansion of the logarithm of the estimated periodogram), which was used in Maharaj and D'Urso (2011), and the symmetric version of the Kulback-Leibler divergence.

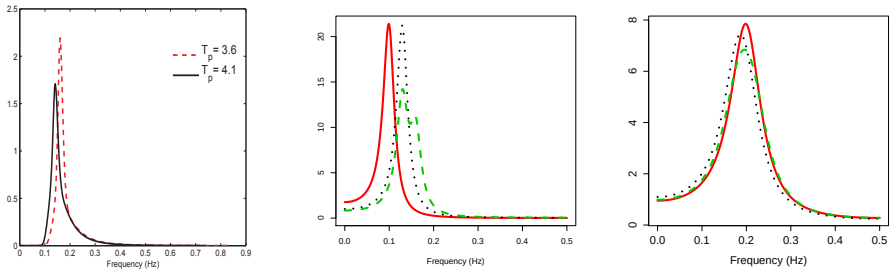
Let $I_X(\omega_k) = T^{-1}\left|\sum_{t=1}^{T} X_t e^{-i\omega_k t}\right|^2$ be the periodogram for time series X, at frequencies $\omega_k = 2\pi k/T,\ k = 1,\ldots,n$ with $n = [(T-1)/2]$, and NI_X be the normalized periodogram, i.e. $NI_X(\lambda_k) = I_X(\omega_k)/\hat{\gamma}_0^X$, with $\hat{\gamma}_0^X$ the sample variance of time series X. The dissimilarity criteria in the frequency domain considered were:

- The Euclidean distance between the normalized estimated periodogram ordinates: $d_{NP}(X,Y) = \frac{1}{n}\Big(\sum_k \big(NI_X(\lambda_k) - NI_Y(\lambda_k)\big)^2\Big)^{1/2}$.
- The Euclidean distance between the logarithms of the normalized estimated periodograms $d_{LNP}(X,Y) = \frac{1}{n}\Big(\sum_k \big(\log NI_X(\lambda_k) - \log NI_Y(\lambda_k)\big)^2\Big)^{1/2}$.
- The square of the Euclidean distance between the cepstral coefficients $d_{CEP}(X,Y) = \sum_{k=0}^{p}\left(\theta_k^X - \theta_k^Y\right)^2$ where, $\theta_0 = \int_0^1 \log I(\lambda)\mathrm{d}\lambda$ and $\theta_k = \int_0^1 \log I(\lambda)\cos(2\pi k\lambda)\mathrm{d}\lambda$.

These dissimilarity measures were compared with the TV distance in a hierarchical algorithm and with the HSM method. In order to compare the distances, the simulation settings were the same in all cases, the estimator of the spectrum is the smoothed periodogram using a Parzen window with bandwidth equal to 100 points, and the clustering algorithm is hierarchical with the complete link function (similar results are obtained with the average link). For the HSM method we denote by $HSM1$ when we use the single version and $HSM2$ when we use the average version.

To evaluate the rate of success, we consider the following index which has been already used for comparing different clustering procedures [Pértega Díaz and Vilar (2010), Gavrilov et al. (2000)]. Let $\{C_1,\ldots,C_g\}$ and $\{G_1,\ldots,G_k\}$, be the set of the g real groups and a k-cluster solution, respectively. Then,

$$\mathrm{Sim}(G,C) = \frac{1}{g}\sum_{i=1}^{g}\max_{1\le j\le k}\mathrm{Sim}(G_j,C_i), \tag{3.4}$$



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Figure 3.6: Spectra used in the simulation study to compare the TV distance with other similarity measures. Each spectrum, with different color and line type, corresponds to a cluster.

where $\text{Sim}(G_j, C_i) = \dfrac{2|G_j \cup C_i|}{|G_j| + |C_i|}$. Note that this similarity measure will return 0 if the two clusterings are completely dissimilar and 1 if they are the same.

In the comparative study, we replicate each simulation setting N times, and compute the rate of success for each one. The mean of this index is shown in Tables 3.2, 3.3 and 3.4 and a box plot of the values obtained is shown in Figures 3.7, 3.8 and 3.9.

We consider three different experiments. The first one is motivated by the applications in Oceanography, where the differences between spectra could be produced by a small change in the modal frequency. The second experiment was designed to test if the proposals are able to distinguish between a unimodal and a bimodal spectrum. Finally, the third one considers models that are frequently used in the study of signals using a frequency domain approach. For all the experiments, the lengths of the time series were $T = 500, 1000$, and 2000.

- **Experiment 1** is based on two different JONSWAP (Joint North-Sea Wave Project) spectra. The spectra considered have significant wave height H_s equal to 3, the first has a peak period T_p of $3.6\sqrt{H_s}$ while for the second $T_p = 4.1\sqrt{H_s}$. Figure 3.6(a) exhibits the JONSWAP spectra, showing that the curves are close to each other. Five series from each spectrum were simulated and $N = 500$ replicates of this experiment were made. In this case the sampling frequency was set

to 1.28 Hz, which is a common value for wave data recorded using sea buoys.

- **Experiment 2** is based on the AR(2) process. Let Z_t^j be the j-th component, $j = 1, 2, 3$, having AR(2) process with $M_j = 1.1$ for all j and peak frequency $\eta_j = .1, .13, .16$ for $j = 1, 2, 3$, respectively. Z_t^j represents a latent signal oscillating at a pre-defined band. Define the observed time series to be a mixture of these latent AR(2) processes.

$$\underset{K\times 1}{\begin{pmatrix} X_t^1 \\ X_t^2 \\ \vdots \\ X_t^K \end{pmatrix}} = \underset{K\times 3}{\begin{pmatrix} \boldsymbol{e}_1^T \\ \boldsymbol{e}_2^T \\ \vdots \\ \boldsymbol{e}_K^T \end{pmatrix}} \underset{3\times 1}{\begin{pmatrix} Z_t^1 \\ Z_t^2 \\ Z_t^3 \end{pmatrix}} + \underset{K\times 1}{\begin{pmatrix} \varepsilon_t^1 \\ \varepsilon_t^2 \\ \vdots \\ \varepsilon_t^K \end{pmatrix}} \tag{3.5}$$

where ε_t^j is Gaussian white noise, X_t^j is a signal with oscillatory behavior generated by the linear combination $\boldsymbol{e}_i^T Z_t^j$ and K is the number of clusters. In this experiment, $K = 3$, with $\boldsymbol{e}_1^T = c(1, 0, 0)$, $\boldsymbol{e}_2^T = c(0, 1, 0)$ and $\boldsymbol{e}_3^T = c(0, 1, 1)$, and the number of draws of each signal X_t^i is 5, Figure 3.6(b) plots the three different spectra. So, we have three clusters with five members each. For this experiment $N = 1000$ replicates were made, and the sampling frequency was set to 1 Hz.

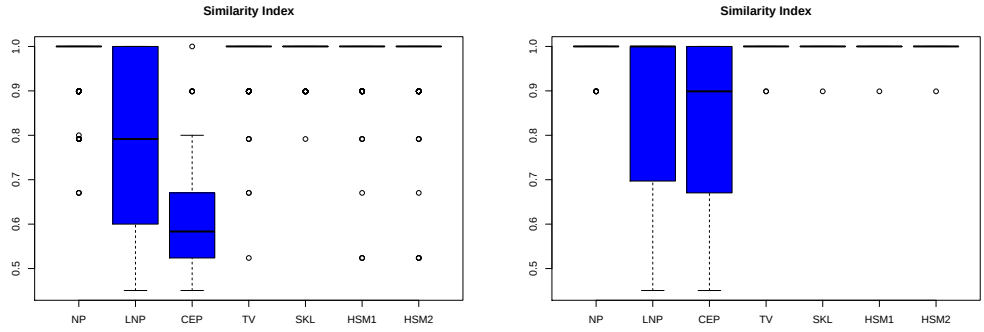
- **Experiment 3** considers time series with two additive components. The first component is an oscillating process with random amplitude while the second one is a random noise with an autoregressive structure. The general model is:

$$X(t) = A\cos(2\pi t\omega_0) + B\sin(2\pi t\omega_0) + Z(t),$$

where $Z(t)$ is an AR(2) process with parameters (η, M), and A, B are independent Gaussian $N(0, 1)$ random variables. We look at three different models, one per cluster.

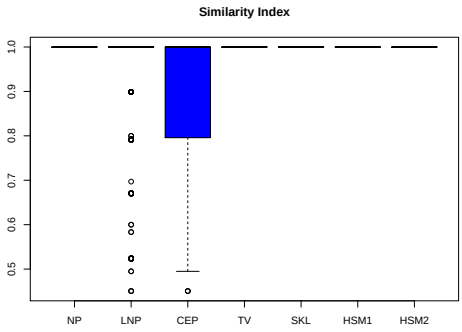
a) Model 1 $\omega_0 = .3$, $\eta = .2$, $M = 1.3$
b) Model 2 $\omega_0 = .1$, $\eta = .18$, $M = 1.3$
c) Model 3 $\omega_0 = .25$, $\eta = .22$, $M = 1.3$

For each model five time series were generated, with $T = 500, 1000, 2000$ and $N = 1000$ replicates of the experiment. Figure 3.6(c) presents the spectral densities for the AR(2) components and shows that they are close to each other. The sampling frequency was set to 1 Hz



(a) $T = 500$, $a = 100$

(b) $T = 1000$, $a = 100$



(c) $T = 2000$, $a = 100$

Figure 3.7: Box plots of the rate of success for the replicates under the simulation setting of Experiment 1, by using different distances.

In all the experiments, we assume the number of clusters known, since the purpose is to compare the dissimilarity measures and not the algorithm to decide the number of clusters.

Tables 3.2, 3.3 and 3.4 give the mean values of the rate of success, for Experiment 1, 2 and 3 respectively. Being a nonlinear function, the logarithm enhances differences when the values of the spectral densities are below 1, and has the opposite effect for values larger than 1. Then, when the spectra are very close and the values are bigger than 1, it is more difficult to distinguish them using a logarithmic distance. This can be seen in the results of **Experiment 1** and **Experiment 3**, where the LNP and CEP distances

Experiment 1

T	a	NP	LNP	CEP	TV	SKL	HSM1	HSM2
500	100	0.979	0.772	0.597	0.988	0.994	0.989	0.988
1000	100	0.998	0.851	0.825	0.999	0.999	0.999	0.999
2000	100	1	0.932	0.908	1	1	1	1

Table 3.2: Mean values of the similarity index obtained using different distances and the two proposed methods in Experiment 1. The number of replicates is $N = 500$.

Experiment 2

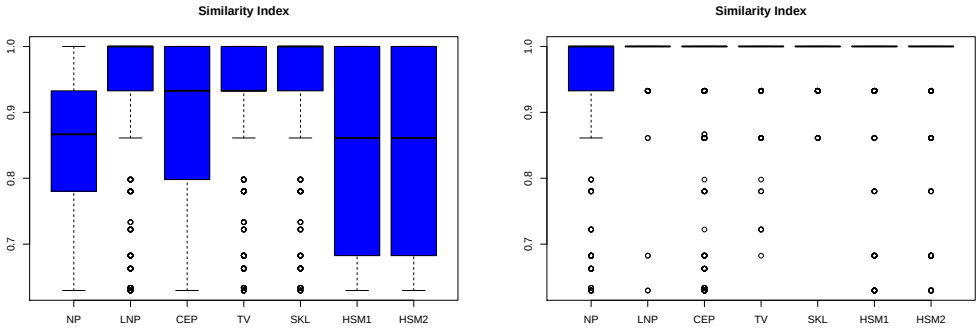
T	a	NP	LNP	CEP	TV	SKL	HSM1	HSM2
500	100	0.864	0.949	0.895	0.930	0.952	0.836	0.838
1000	100	0.961	0.996	0.974	0.990	0.994	0.983	0.983
2000	100	0.995	1	0.999	0.999	0.999	0.999	0.999

Table 3.3: Mean values of the similarity index obtained using different distances and the two proposed methods in Experiment 2. The number of replicates is $N = 1000$.

Experiment 3

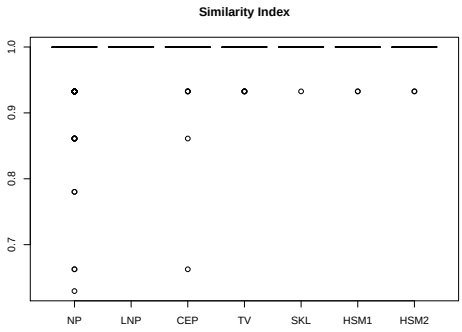
T	a	NP	LNP	CEP	TV	SKL	HSM1	HSM2
500	100	0.974	0.843	0.777	0.976	0.984	0.977	0.977
1000	100	0.974	0.830	0.762	0.975	0.985	0.977	0.976
2000	100	0.975	0.823	0.757	0.975	0.984	0.977	0.977

Table 3.4: Mean values of the similarity index obtained using different distances and the two proposed methods in Experiment 3. The number of replicates is $N = 500$.



(a) $T = 500$, $a = 100$

(b) $T = 1000$, $a = 100$

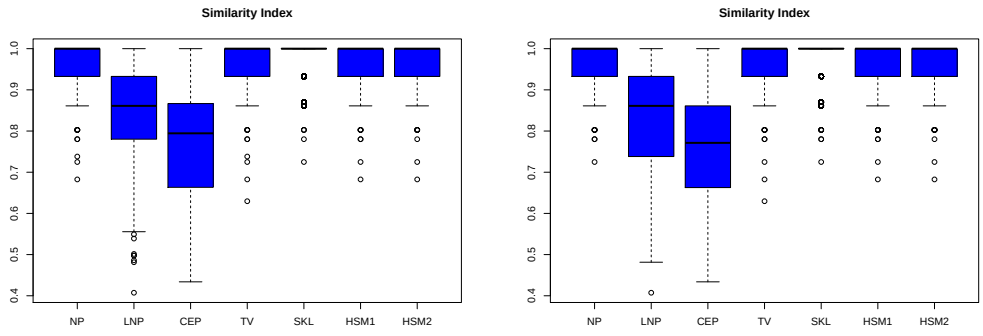


(c) $T = 2000$, $a = 100$

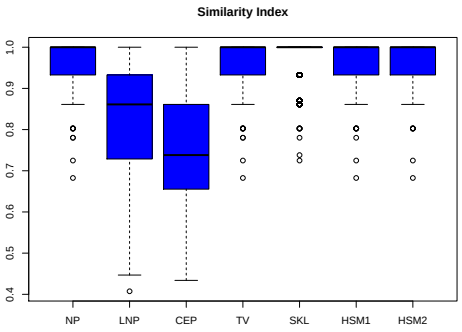
Figure 3.8: Box plots of the rate of success for the replicates under the simulation setting of Experiment 2, by using different distances

have smaller rate of success compared to the TV distance. In **Experiment 1**, for short series ($T = 500$) the best results correspond to SKL followed closely by the TV distance, while for long series $T = 1000$, and 2000 the methods that used the TV distance have a success rate close to one.

In **Experiment 2** and **Experiment 3** it is more difficult to discriminate between groups, since there are three clusters involved. In **Experiment 2** the best results were obtained with the SKL distance, followed by the LNP, while the TV distance was not far behind. In **Experiment 3**, the best results were obtained with the SKL divergence, and the methods based on the TV distance were very close.



(a) $T = 500$, $a = 100$ (b) $T = 1000$, $a = 100$



(c) $T = 2000$, $a = 100$

Figure 3.9: Box plots of the rate of success for the replicates under the simulation setting of Experiment 3, by using different distances

Figures 3.7, 3.8, and 3.9 show the boxplot for the values of the rate of success obtained in each experiment. We can see from the box plots corresponding to **Experiment 1** that the CEP distance has many values smaller than 0.9 even in the case of $T = 2000$.

In **Experiment 2**, the HSM method did not have a good performance, in small and medium sized time series, compared to the others. It is necessary to have $T = 2000$, for the HSM method to identify the clusters more precisely. The NP distance has the worst performance overall. In **Experiment 3**, the performance of all distances does not improve significantly when we increases the length of the series. The LNP and CEP distances instead of improving, get worse when we increases the value of T.

In general, the rate of success for the methods that used the TV distance are good, in some cases they have the best results, and when they do not, they are close to the best. The methods based on logarithms, such as the LNP and CEP, have in some cases a good performance but in others their performance is very poor. The SKL has the best results in many cases, however, as we mentioned in Chapter 2, this distance cannot be computed when two spectra have disjoint support. In addition, methods that use logarithmic functions require more computational time than methods based directly on spectra.

It is important to mention that these methods could be applied to big data, long time series or several series. In general, the proposed methods are efficient in this sense. The computational complexity is $O(n^3T)$, where n is the number of time series to be clustered and T the length of each time series. It implies that the computational time does not increases exponentialy as with other methods.

Considering the properties of the TV distance and its performance when used as a dissimilarity measure in a clustering method, we consider the procedures proposed as a good option as time series clustering methods.

3.4 Detection of transitions between spectra

In many instances, non-stationary time series present changes that do not occur abruptly, but rather appear as transitions between stationary intervals. In such cases, methods devised for the detection of changes usually produce poor results. One example that will be considered in detail in the next chapter, is the analysis of stationary periods for wave height data. The sea surface is stationary only for a period of time, and when the environmental

conditions that produce it change, there is usually a slow transition, lasting hours or days, to a new stationary state.

As an alternative to change-point detection, we propose looking at the behavior of segments of time from a long time series, and using clustering algorithms in order to detect periods having similar behavior. If these periods are contiguous in time, then it is natural to consider them as part of a longer stationary interval. This section is devoted to presenting a simulation method for transitions, which is later used in a simulation experiment to evaluate the performance of the procedure.

3.4.1 Simulation of transitions between two spectra

We propose a new method to simulate transitions between two stationary periods of a time series. Our proposal is based on the definition of a Locally Stationary Process.

From Definition 1.2, $X_{t,T}$ has a unique time varying spectral density which is locally the same as the spectral density of $\tilde{X}_t(u)$. Furthermore, it has locally the same auto-covariance since $cov(X_{[uT],T}, X_{[uT]+k,T}) = c(u,k) + O(T^{-1})$ uniformly in u and k, where $c(u,k)$ is the covariance function of $\tilde{X}_t(u)$. This justifies taking $c(u,k)$ as the local covariance function of $X_{t,T}$ at time $u = t/T$.

We need a method that, given two different spectra, is able to simulate a process that changes its spectrum from f_1 to f_2 during the transition period. This can be reformulated equivalently in terms of the covariance functions r_1 and r_2.

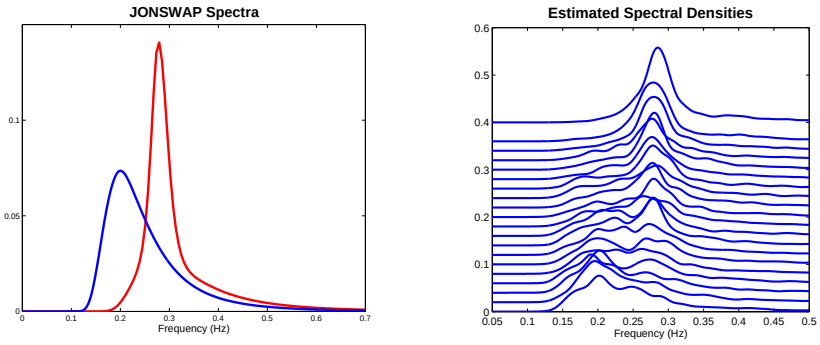
Suppose that we have two independent processes X_t^1 and X_t^2, which have $r_1(h)$ and $r_2(h)$ as covariance functions. Take

$$X_t = \sqrt{a(t)}X_t^1 + \sqrt{b(t)}X_t^2,$$

where a and b are functions with slow changes, $a(t) \approx a(t+h)$ and $b(t) \approx b(t+h)$ if h is small, $a(0) = b(T) = 1$ and $a(T) = b(0) = 0$. Then it is easy to see that, for small values of h,

$$\begin{aligned}\operatorname{cov}(X_t, X_{t+h}) &= \sqrt{a(t)}\sqrt{a(t+h)}r_1(h) + \sqrt{b(t)}\sqrt{b(t+h)}r_2(h) \\ &\approx a(t)r_1(h) + b(t)r_2(h).\end{aligned}$$

So, X_t is a process that for t near 0 has locally covariance function r_1 and for t near T it has r_2.



(a) (b)

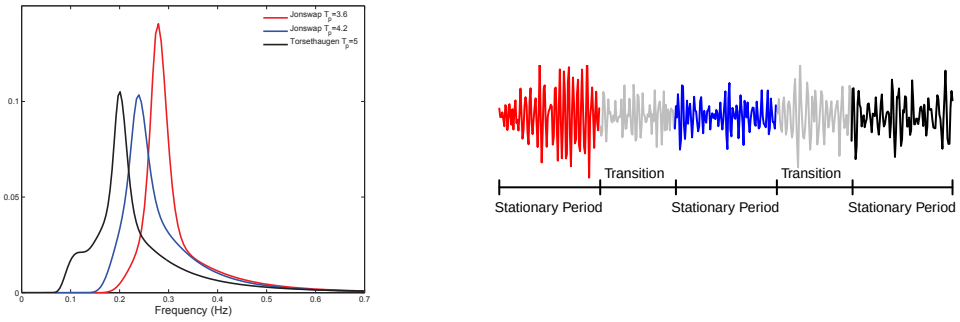
Figure 3.10: Simulation of a transition between two spectral densities. (a) The transition starts with $T_p = 5$ (the blue curve) and finishes with $T_p = 3.6$ (the red curve). (b) Estimated densities from the simulated data during the transition.

Example 3.3. To test our method we have to produce data from a process that has a transition period, with a slow change from one spectrum to another. Take f_1 and f_2 as JONSWAP spectra with $H_s = 1$ in both cases and $T_p = 5$ and 3.6 respectively. We choose $a(t) = 1 - T/t$ and $b(t) = 1 - a(t)$, where $T = 5$ hours is the total observation time. Figure 3.10(a) shows the spectra involved in the transition, starting with the blue spectrum and finishing with the red spectrum. Figure 3.10(b) shows the estimated densities after we apply the algorithm, we can observe the form of the transition and how the process starts at f_1 and finishes at f_2.

3.4.2 Detection of transitions

Further simulation studies were carried out to assess the performance of the clustering algorithm in the presence of transition periods. The main objective was to gauge the performance when slow transitions between stationary periods are present in a data set.

Experiment 4. The simulations were carried out using the JONSWAP and Torsethaugen families of spectra (see Torsethaugen, 1993; Torsethaugen and Haver, 2004). The latter is a family of bimodal spectra used in Oceanography, which accounts for the presence of swell and wind-generated waves, and was also developed to model spectra observed in North-Sea locations. In all cases, the significant wave height (H_s) was set to 1. The



(a) (b)

Figure 3.11: Elements of Experiment 4. (a) Spectra involved in the stationary periods. (b) Sketch of the simulation sequence (100 points per period).

simulation has three stationary periods and two transitions between them, both transitions lasts 3 hours from one stationary period to the next.
Stationary Period 1 - the simulated series starts with waves from a stationary period of 4 hours, from a JONSWAP spectrum with peak period $T_p = 3.6$, *Stationary Period 2* - the second stationary period corresponds to another 4 hours of waves drawn from a JONSWAP spectrum with $T_p = 4.2$, and *Stationary Period 3* - a third 4-hour stationary period but in this case from a bimodal family, Torsethaugen spectrum with $T_p = 5.0$.
In this case, we simulate $N = 1000$ replicates and the sampling frequency was set to 1.28 Hz.

Figure 3.11(b) shows a sketch of the simulation setting, where we get one continuous signal. We start with the stationary period in red which corresponds to the red spectrum in Figure 3.11(a), then a transition period in gray color, and so on. Figure 3.11(a) plots the spectra involved in the experiment for the stationary periods.
The test procedure is the following.

1. Each time series has 82944 time points, 4608 points per hour (18 hrs).

2. Data are divided into 30-minute intervals, each segment will be considered as an element to be clustered.

3. Then, we have 36 segments, each of length $T = 2304$. We apply the clustering methods to these segments.

First, we consider that there are just three genuine clusters, since the transition periods, by definition, do not represent intervals with a homogeneous behavior. We consider the TV distance in a hierarchical algorithm with the complete link function and the HSM with the two possible algorithms, $HSM1$ and $HSM2$.

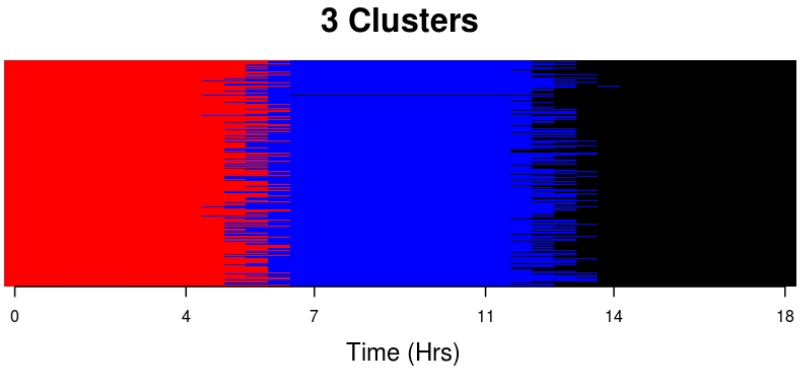
Figure 3.12 shows the results obtained if we set the number of clusters to 3. Each plot represents by the corresponding color (red, blue and black) the members that are assigned to the same cluster. For the three methods, the resulting clusters contain each one of the stationary periods, 0-4 hours, 7-11 hours and 14-18 hours. The beginning of the transitions are mostly assigned to the previous stationary period, for example the two segments from 4 to 5 hours are assigned to the same cluster as the first stationary period. While the end of the transitions are assigned to the next stationary period. This is reasonable since these are the more similar periods, respectively. The middle of the transitions seem to be assigned randomly between the two closest stationary periods.

It is interesting to observe that, in general, the elements in a cluster are contiguous in time, even when no information about the time structure of the series is included in the procedure, and the methods identify the changes in the transition periods.

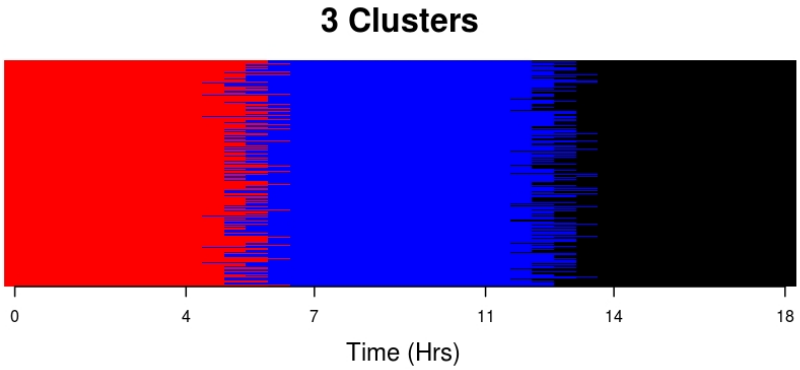
The problem of deciding whether the intervals close to the border belong to a cluster or should be classified as transition periods requires a criterion for deciding whether a given interval is "well classified" within a given cluster. In Alvarez-Esteban et al. (2016a), we explore the use of the silhouette index, proposed by Rousseeuw (1987), which gives a measure of the adequacy of each point to its cluster.

Another approach that was also attempted was the use of trimming procedures in the clustering process, as is considered in the work of Cuesta-Albertos and Fraiman (2007) for functional data. In this context, the spectral densities would be the functional data to be classified. The trimming procedure "discards" a certain fraction of the information in the classification process, in order to robustify the result, and it seems reasonable to consider the trimmed information as data objects that do not fit properly within any of the clusters. In consequence they could be labelled as transition periods. An important shortcomming of this method is the long time it takes even with moderately sized samples, and therefore the difficulty in handling real-life information.

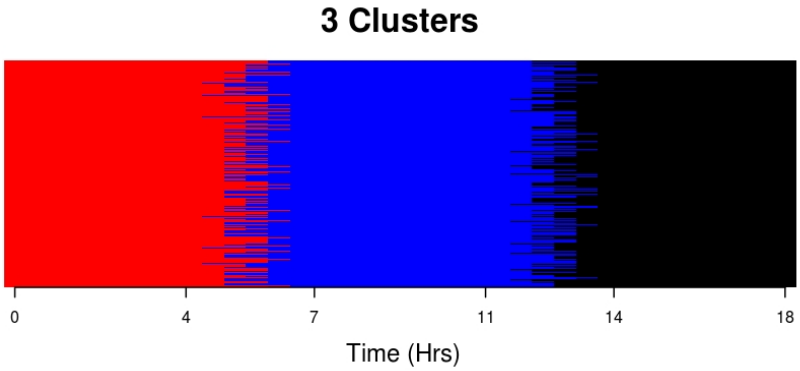
As an alternative one could consider that there should be five clusters,



(a) TV distance in a hierarchical algorithm (complete linkage).

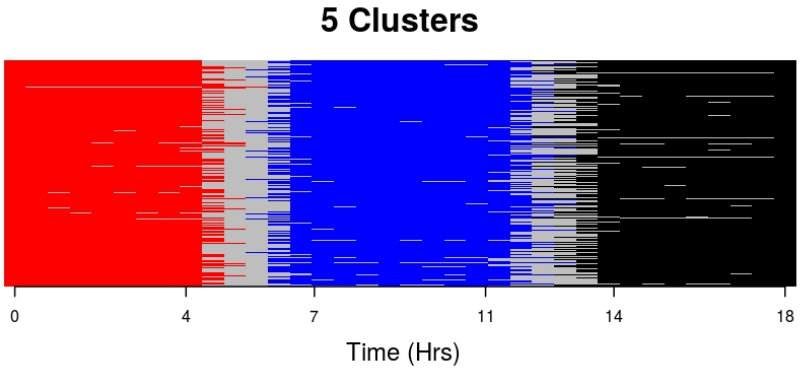


(b) HSM method with the single version.

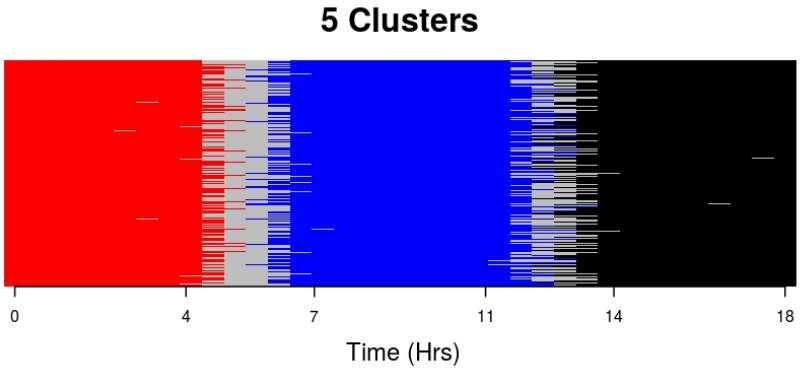


(c) HSM method with the average version.

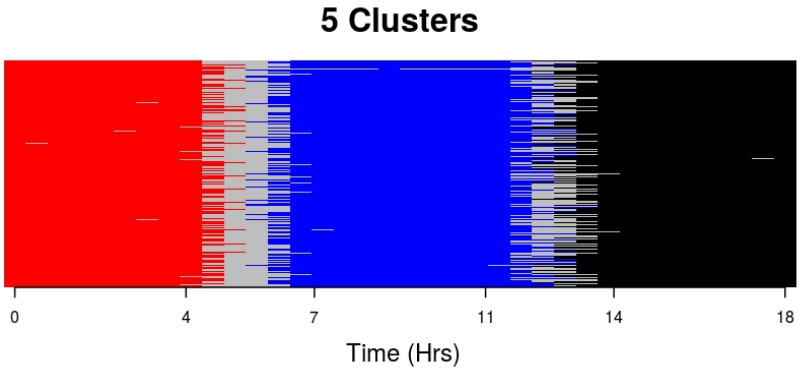
Figure 3.12: Members of groups 1 (red), 2 (blue), and 3 (black). If we consider only 3 clusters.



(a) TV distance in a hierarchical algorithm (complete linkage).



(b) HSM method with the single version.



(c) HSM method with the average version.

Figure 3.13: Members of groups 1 (red), 3 (blue), and 5 (black), which correspond to each one of the stationary periods.

three corresponding to the stationary periods and two to the transitions. However, the transitions can be difficult to identify as a cluster, since there is not complete homogeneity between elements of a transition.

Figure 3.13 shows the results with 5 clusters. Each plot represents by the corresponding color (red, blue and black) the clusters that are assigned to the stationary periods. In gray we represent the two clusters that should correspond to transitions. As we can see, none of the different methods are able to recover the complete transitions. However, the results are reasonable, in the sense that the beginning of the transitions are assigned to the previous stationary period and the end of the transitions to the next stationary period.

The transition from a unimodal to a bimodal spectrum (transition 2) is more difficult to identify. The hierarchical method with the TV distance has a better performance with the transitions, keeping the two central segments of a transition in the same cluster more often than the HSM method. It seems, that the HSM method has more difficulty in identifying slow transitions.

In practice, it is important to consider that the beginning and end of transitions are going to be hard to identify if the transition is too slow.

3.5 Unknown number of clusters

An important point to discuss is how to choose the number of clusters. There is not a universally best criterion that works for all clustering methods. Many researches are still working on improving the existing criteria or in proposing new ones. Usually the criteria will not be fully automatic and will depend on the problem.

The general theory of clustering has some options to decide the number of clusters. The criteria are usually of two types, internal and external. With an external criterion, we need to have some prior information of the true clusters. An internal criterion is usually more reasonable, since in real data analysis we do not always have such information.

Dunn's Index. Among those indices that admit as input a dissimilarity matrix, we have selected Dunn's index. This index is defined as

$$V_D(k) = \min_{1\leq i\leq k}\left\{\min_{i+1\leq j\leq k}\left(\frac{D(C_i,C_j)}{\max_{1\leq h\leq k} diam(C_h)}\right)\right\},$$

where k is the number of clusters, $D(C_i, C_j)$ is the distance between clusters C_i and C_j, and $diam(C_h)$ is the diameter of cluster C_h.

From the definition of V_D it is clear that high values point to suitable values of k. However, the maximum value of $V_D(k)$ is not always the best choice, specially when we have patterns which include clusters which are close to each other. This situation is common in random sea waves where consecutive stationary periods can have similar characteristics. All these issues make the choice of the "optimal" k not an automatic process.
Since this is a well known criterion, we will not present any simulation in this case. In applications, the results obtained with this index are similar to other indices such as the David-Bouldin Index.

Test based on the distribution of $\hat{d}_{TV}$. Another procedure for deciding the number of clusters is based on the bootstrap algorithm proposed in Chapter 3. We will use this methodology to approximate the distribution of the total variation distance between two clusters. Note that due to the hierarchical structure of the algorithms used in all the methods proposed, the test

$$H_0: \ k-1 \text{ Clusters} \qquad \text{vs} \qquad H_A: \ k \text{ Clusters},$$

is equivalent to the test,

$$H_0: \ 1 \text{ Cluster} \qquad \text{vs} \qquad H_A: 2 \text{ Clusters}.$$

This is because the $(k-1)$ clusters are built by joining two of the k clusters.

The distribution of the total variation distance between two clusters depends on the clustering procedure. When using the HSM method we aim to approximate the distribution of the distance between the mean spectra in each cluster while for the hierarchical clustering with the TV distance, we need to produce samples from each cluster to approximate the distribution of the distance calculated through the link function.

The procedure of this test will be:

- Run the clustering procedure, either the HSM method or hierarchical clustering with the average or complete linkage.
- Identify the two clusters that are joined to get the $(k-1)$ clusters.
- Consider as the estimation of the common spectra, $\hat{f}$, the mean spectra over all elements in both clusters.
- Simulate with the bootstrap procedure the spectra to compute the TV distance. We consider two cases:

Experiment 1				
Test	α	Complete	Average	HSM
1 cluster vs 2 clusters	0.01	1	1	1
	0.05	1	1	1
	0.1	1	1	1
2 clusters vs 3 clusters	0.01	0.052	0.154	0.008
	0.05	0.206	0.492	0.058
	0.1	0.382	0.670	0.164

Table 3.5: Proportion of times that the null hypothesis is rejected. Complete corresponds to the TV distance in hierarchical algorithm with the complete link function, and Average with the average link.

Case 1. When using the HSM method simulate two spectral densities from the common spectra f and compute the TV distance between them. We repeat this procedure M times.method.

Case 2. When using hierarchical clustering with the TV distance simulate two sets of spectral densities of size g_1 and g_2 from the common spectra f, where g_i are the number of members in cluster $i = 1, 2$ (clusters to be joined). We compute the link function (complete or average) between these two sets of spectra using the TV distance.

- Run the test proposed in Chapter 3 with the bootstrap sample.

Remark. Notice that this test assumes that there exits a common spectra f.

To explore the performance of our proposals, we used **Experiments 1** and **2**. We consider the TV distance to feed a hierarchical algorithm with two different link functions, average and complete. Also, we consider the HSM method with the hierarchical spectral merger algorithm. In this case, we just use 500 replicates for each experiment.

Tables 3.5 and 3.6 present the proportion of times that the null hypothesis is rejected. To reject we consider a bootstrap quantile of probability α. We do not expect to have a proportion of rejection equal to α, since in the case of using the complete or average link, these values are not a direct observation of the TV distance. However, we expect to have a good performance. In general, it could be possible to overestimate the number of clusters.

In **Experiment 1** the true number of clusters is 2. From Table 3.5, we observe that all methods reject the hypothesis of one cluster, at all the

Experiment 2

Test	α	Complete	Average	HSM
2 clusters vs 3 clusters	0.01	0.968	1	0.25
	0.05	1	1	0.924
	0.1	1	1	0.998
3 clusters vs 4 clusters	0.01	0.072	0.18	0.002
	0.05	0.228	0.924	0.050
	0.1	0.376	0.998	0.106

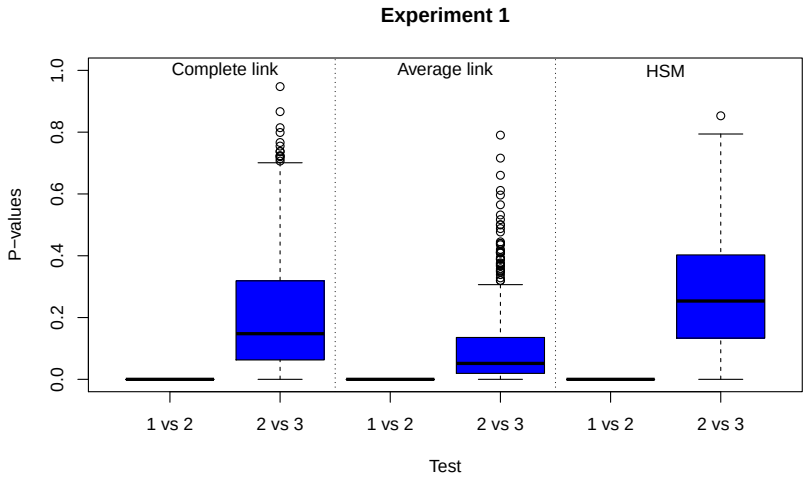
Table 3.6: Proportion of times that the null hypothesis is rejected, in Experiment 2.

significance levels. This means that the procedure will not under estimate the number of clusters. To test 2 vs 3, the proportion of rejection is high when we use the average link function, except in the case of $\alpha = 0.01$. If we use the complete link, the results are better. However, the best results are for the HSM method.

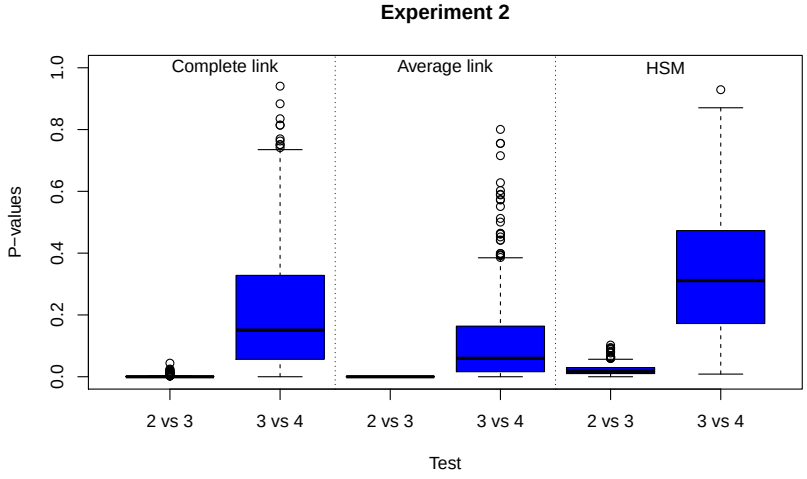
In **Experiment 2** the true number of clusters is 3. This is a more a difficult case, since the spectra are very close. From Table 3.6 when testing 2 vs 3 clusters, we observe that the complete and average link functions do not under estimate the number of clusters. However, the HSM method can not distinguish 3 clusters at a level $\alpha = 0.01$, but it is possible at higher levels. For testing 3 vs 4 clusters, the performance for the HSM and complete are better, being again necessary a small value of α for the average link to have reasonable performance.

Figure 3.14 shows the p-values obtained comparing the value obtained at each simulation with the bootstrap distribution. We confirm the fact that the underestimation of the number of clusters has low probability, almost zero in some cases, for the three methods. When the number of clusters to test is the correct one, 2 in **Experiment 1** and 3 in **Experiment 2**, the p-values are widely distributed in the case of the complete link and HSM method. With the average link, the p-values are smaller compared to the other methods. In general, this test has a good performance when one uses the complete link or the HSM method.

Permutation Test. Consider the same testing problem, as an alternative for the average link function, we can use a permutation test (see Rudolph, 1995).



(a)



(b)

Figure 3.14: P-values obtained in the test of number of clusters using bootstrap samples.

Experiment 1			Experiment 2		
Test	α	Average	Test	α	Average
1 cluster vs 2 clusters	0.01	.848	2 clusters vs 3 clusters	0.01	.848
	0.05	1		0.05	.994
	0.1	1		0.1	.996
2 clusters vs 3 clusters	0.01	.002	3 clusters vs 4 clusters	0.01	.022
	0.05	.006		0.05	.032
	0.1	.182		0.1	.218

Table 3.7: Proportion of times that the null hypothesis is rejected, using the permutation test, in Experiment 1 and 2.

Let be $G_1 = \{f_1^1, f_2^1, \ldots, f_{n_1}^1\}$ and $G_2 = \{f_1^2, f_2^2, \ldots, f_{n_2}^2\}$ two clusters where $f_{i_j}^j$ are the spectral densities of the time series $X_{i_j}^j$, members of the clusters G_j, $j = 1, 2$ and $i_j = 1, \ldots, n_j$. If the two clusters belong to a one bigger cluster, $G_1 \cup G_2 \subseteq G$, we can take a subsample of this clusters as follows.

- Let be $G^* = G_1 \cup G_2 = \{f_1^1, f_2^1, \ldots, f_{n_1}^1, f_1^2, f_2^2, \ldots, f_{n_2}^2\}$.
- Take, with probability $\frac{1}{n_1+n_2}$, n_1 elements of G^* and assign them to G_1^*.
- Take $G_2^* = G^* \setminus G_1^*$.
- Finally, compute the average link function between G_1^* and G_2^*.
- Repeat this procedure M times.

Then, the permutation test will take the link functions values computed between the subsample clusters as a sample of the distribution of our statistic. So, the test will reject the null hypothesis using the quantiles of this sample.

Table 3.7 shows, for both experiments, the proportion of times that we reject the null hypothesis using the average link function and the permutation test. We observe that the level of the test is improved, however, it loses some power.

Remark. This test is not useful when the complete link function is used. Due to the hierarchical algorithm, the maximum between the original groups will always be bigger or equal than the maximum of any of the subsamples.

3.6 Discussion

The use of the TV distance as a dissimilarity measure for clustering has shown good results compared to other dissimilarity measures proposed in the literature. In some of the experiments in the simulation study we got the best rate of success or close to the best ones. In addition, the clusters generated by our proposal have an intuitive interpretation in terms of real application problems. In the case of transitions, it is still difficult to identify the beginning or end of the transitions, however, the results are acceptable and a good approximation of the true clusters, if we consider a transition as a cluster. The HSM method seems not to be a good option for the detection of transitions.

The election of the number of clusters will always be complicated. However, the test proposed is a promising option. In particular, the HSM method has a good performance using this test to choose the number of clusters.

We proposed the use of time series clustering methods to detect changes in spectra. However, the resolution of the change point detected will depend on the time series length, since we need a reasonable number of time points to have a good estimation of the spectral density.

The proposed methods are general for time series clustering, they can be used to identify similarities in time and/or space. Evenmore, they can be used to cluster any set of time series where the goal is to find similarities in spectra. These methods were proposed in Alvarez-Esteban et al. (2016b); Euán et al. (2015). Further details and discussion can be found in them.

Chapter 4

Applications to Data

In this chapter we present the analysis of two different applications, both of which are commonly studied using the spectral analysis of time series. The spectral density has many interpretations in each case. First, we consider an application to the analysis of ocean wave data and then we present the analysis of brain signals.

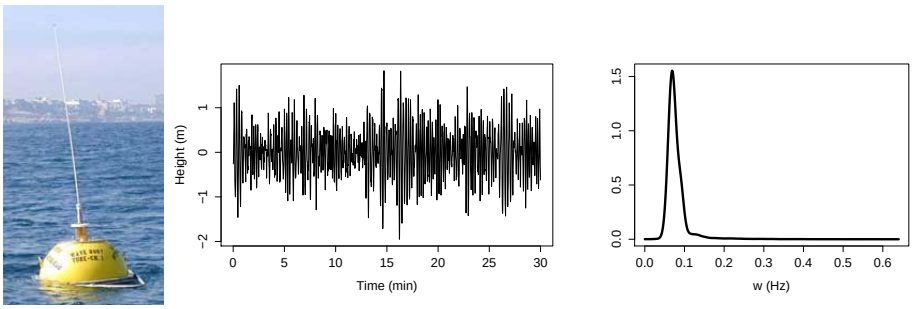
4.1 Ocean wave analysis

Random processes have been used to model sea waves since the 1950's, starting with the work of Pierson (1955) and Longuet-Higgins (1957). Models based on random processes have proved useful, allowing the study of many wave features (see, e.g. Ochi, 1998). A class of models often used to study sea waves in deep waters with standard conditions are stationary centered Gaussian processes (Aage et al., 1999; Ochi, 1998). The intuitive idea was to consider a linear model based on the superposition of infinite elementary waves of the form

$$\zeta_n = Re(A_n e^{i(\lambda_n x + \mu_n y + \omega_n t)}),$$

where A_n is a centered Gaussian variable, Re is the real part and (x, y) is a specific location.

The stationarity hypothesis allows the use of Fourier spectral analysis to study the wave energy distribution as a function of frequency. In particular, this spectral analysis is related to several features of interest, such as the significant wave height (H_s) or the dominant or peak period (T_p), that can be computed from the spectral distribution (see, e.g. Ochi, 1998).



(a) Buoy 106 (b) 30 minute wave (c) Estimated Spectrum

Figure 4.1: One interval of the data set taken by Buoy 106. (a) Buoy at Waimea Bay, Hawaii. (b) A 30 minute wave (centered) taken at 1.28 Hz. (c) Estimated spectrum of the wave process.

Gaussian models, beyond being a good first order approximation, allow obtaining explicit expressions for the distribution of objects of interest. An accurate description of the statistical characteristics of the wave climate on a given region is an important input for the design of marine structures and ships and also for the design of wave energy converters.

However, the property of stationarity is only true for short time intervals and frequently the changes are not abrupt. These changes could be considered as a transition between stationary periods. A natural question is, how long can we consider a sea state to be stationary? How long can a transition last? So, we are interested in detecting stationary and transition periods.

Typically, stationary sea states last for some time (hours or days), and then, due to changing weather conditions, sea currents, the presence of swell or other reasons, change to a different state. The idea of our analysis is to identify short stationary intervals which have similar behavior, in terms of their spectral densities. If these intervals are contiguous in time, then it is reasonable to assume that they constitute a single (longer) stationary interval.

4.1.1 Data description

The data at a fixed point (x, y) is recorded by buoys that are located in the ocean. Usually, data are sampled at a frequency of 1.28 Hz which is 5 data points per 4 seconds approximately.

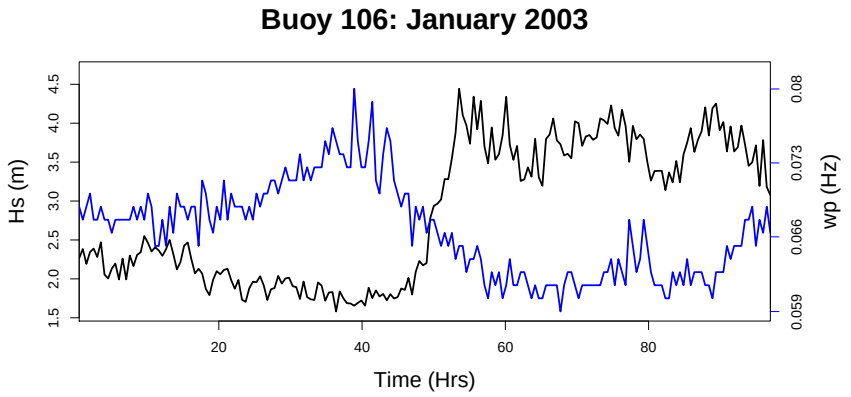


Figure 4.2: Buoy 106, significant wave height and modal frequency of each segment.

We use raw wave height time series obtained from the U.S. Coastal Data Information Program (CDIP) website. The data considered were measured by a moored buoy: Buoy number 106 (number 51201 for the National Data Buoy Center) which is located in Waimea Bay, Hawaii, with a water depth of 200 m. We consider the month of January, 2003. Figure 4.1 shows the wave height time series corresponding to a 30 minute interval and the estimated spectral density.

4.1.2 Results using the TV distance as a similarity measure

The complete recorded data set considered corresponds to 92 hours. In Oceanography it is usual to divide long wave height records into shorter intervals of between 20 to 30 minutes, and then calculate the spectral density. These intervals are considered to be short enough for the stationarity assumption to hold, yet long enough to have a reasonably accurate estimation of the spectrum. Based on these estimated spectra, we can compute the significant wave height and modal frequency of each spectra, which will serve as a summary of the data.

The significant wave height (H_s) is the mean wave height of the highest third of the waves, under the Gaussian assumption it is equal to

$$4\sqrt{m_0},$$

where $m_0 = \int_{-\infty}^{\infty} f(\omega)\mathrm{d}\omega$.

The modal frequency, ω_p, is the frequency at which a wave spectrum reaches its maximum, the inverse of the peak period. Figure 4.2 shows H_s (black line) and $\omega_p = \frac{1}{T_p}$ (blue line) computed for each time segment. From this plot we have a description about the behavior of the waves recorded by Buoy 106. During the first 50 hours the significant wave height stays mainly below 2.5 m. and for a long interval it is below 2 m. Around 50 hours it rises to about 4.5 m. and stays above 3 m. for the rest of the period. The dominant frequency decreases from a 0.073 to 0.06 Hz, approximately.

Our goal is to find the stationary intervals and also look at the changes in spectra between the different intervals. The procedure to analyze the data is the following:

- Divide the data into segments of 30 minutes, 2304 time points.
- Each segment is considered as a unit, so we apply the clustering procedure to 192 "time series" (each one is one segment).
- We use the TV distance between the smoothed estimated spectra (Parzen window with a bandwidth $a = 100$) to feed a hierarchical clustering algorithm with a complete or average link function.
- If two consecutive (in time) segments are in the same cluster, then they will be considered to be part of a stationary period.

We get two different (but similar) results, one with the complete link and the other with the average link. Figure 4.3 shows the value of Dunn's index in each case. The "best" number of clusters should be the one where Dunn's index reaches its maximum, however, the maximum value is not always the best choice. The two highest values were considered and, after analyzing the results it was observed that in most cases, the second highest value gave the best clustering (see Alvarez-Esteban et al., 2016b). So, for the complete link function we choose 6 clusters and for the average link function we choose 5 clusters. Figures 4.4 and 4.5 show the dendrograms resulting in each case and the resulting branches if we cut the tree at 6 and 5 clusters respectively.

As in the simulation study (**Experiment 4** in **Section** 3.4.2), most of the members in a cluster are contiguous segments in time, even though the time structure plays no role in the clustering algorithm.

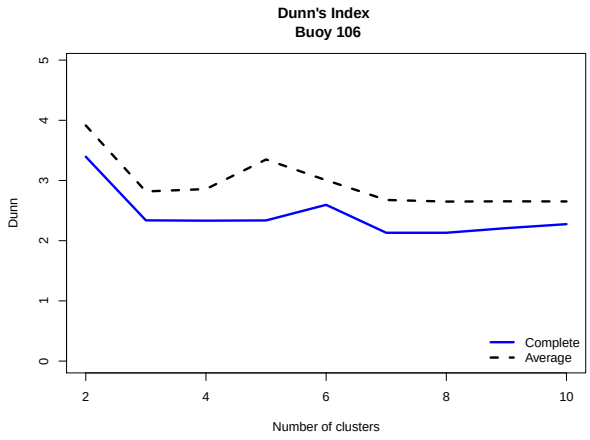


Figure 4.3: Dunn's Index computed between 2 to 10 clusters for the complete and average link functions.

A shortcoming of hierarchical clustering algorithms is that, once an element has been assigned to a cluster, it cannot be reassigned to a different one, even if changes in the composition of the clusters indicate that it would have been better classified on a different cluster.

The silhouette index, proposed by Rousseeuw (1987), gives a measure of the adequacy of each point to its cluster. Let $a(i)$ be the average distance or dissimilarity of point i with all the other elements within the same cluster, and let $b(i)$ be the smallest average dissimilarity of i to any of the clusters to which i does not belong. Then the silhouette index of i is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

This index satisfies $-1 \leq s(i) \leq 1$ for all i, and large positive values indicate that the element has been well classified while negative values point to misclassification. As a consequence the classification of intervals with negative silhouette index was revised. Just a few number of segments were reasigned.

Figures 4.6 and 4.7 show the results of the clustering procedure for the average and complete linkage functions after the correction using the silhouette index, respectively. In part (a), we show which segments belong to each cluster with vertical lines in different colors, i.e, each color represents one cluster and time segments with vertical lines in the same color are members

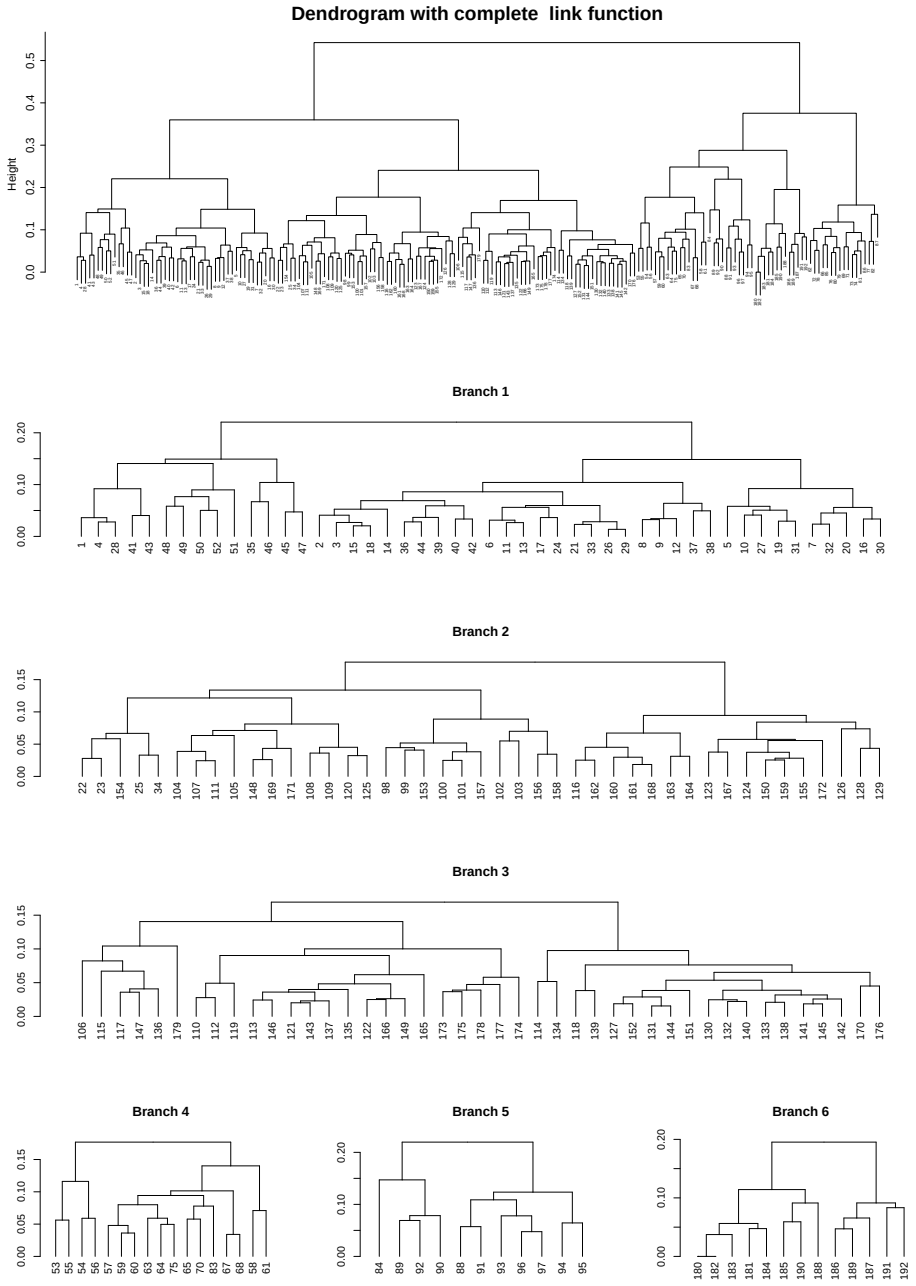


Figure 4.4: Dendogram of Buoy 106 using the complete link function, the index is the number of the segment (each segment is a 30 minute recording). Top: Complete Dendogram. Bottom: Low branches when we cut the dendrogram at 6 clusters.

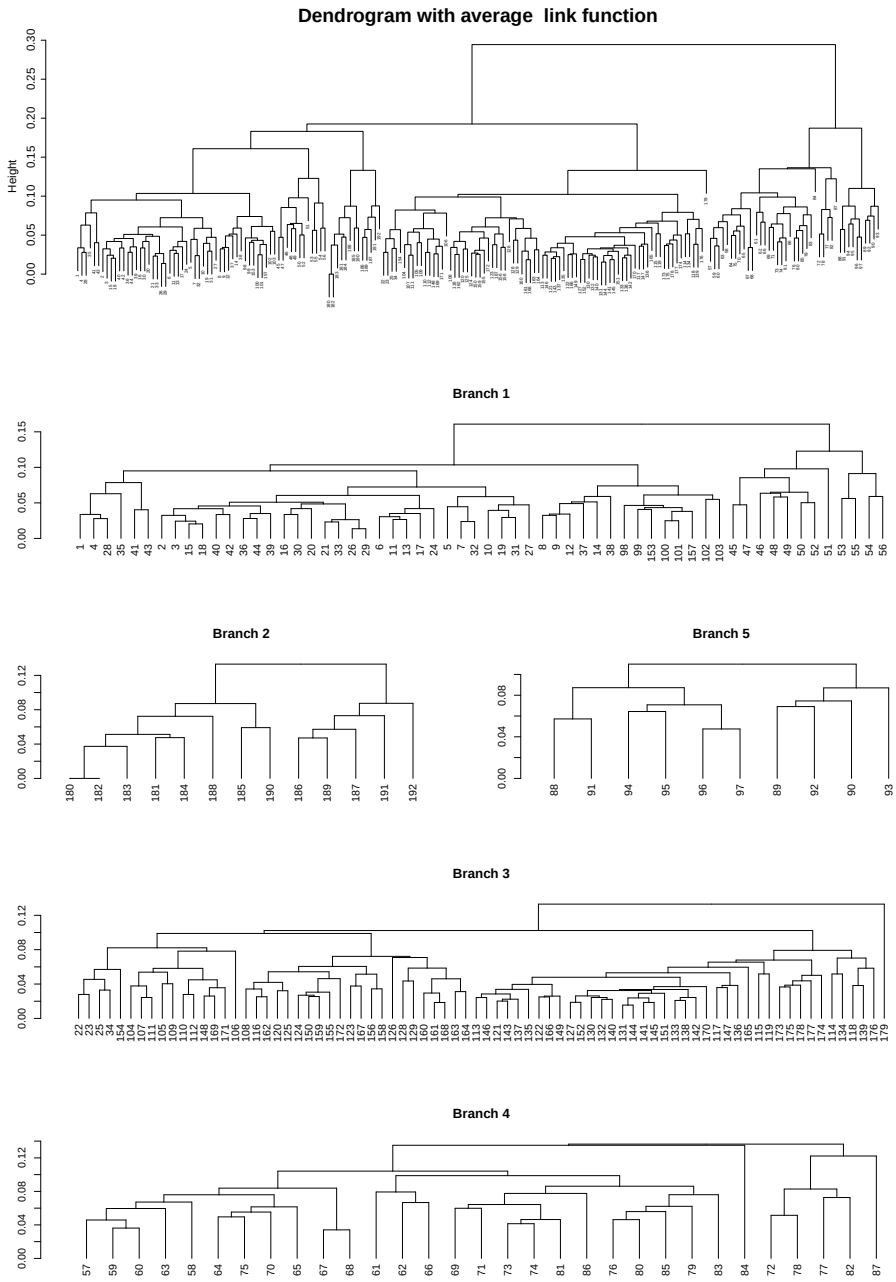


Figure 4.5: Dendrogram of Buoy 106 using the average link function, the index is the number of the segment (each segment is a 30 minute recording). Top: Complete dendrogram. Bottom: Low branches when we cut the dendrogram at 5 clusters.

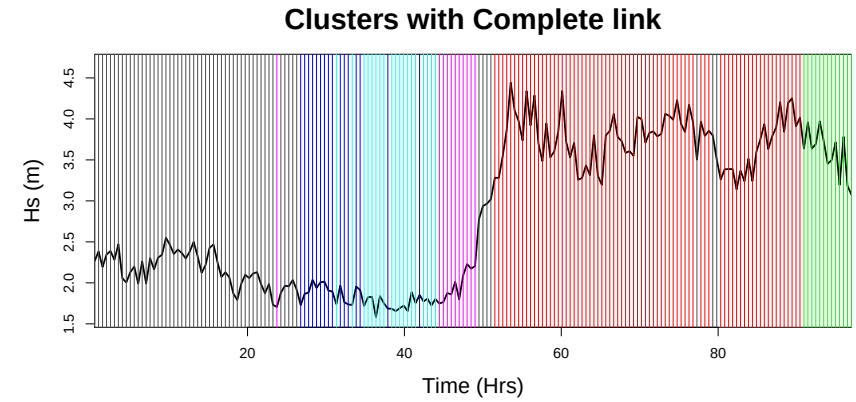
of the same cluster. In part (b), we show (with the corresponding color) the estimated spectra of all members in a cluster and in black the mean spectrum.

As we mentioned before, the clustering procedure captures the time structure in the data using only information about the TV distance between normalized spectral densities. In addition, using either the complete or average link, the members in a cluster have very similar spectra and the method is able to identify small differences between clusters. For example, the method is able to discriminate between unimodal and bimodal spectra.

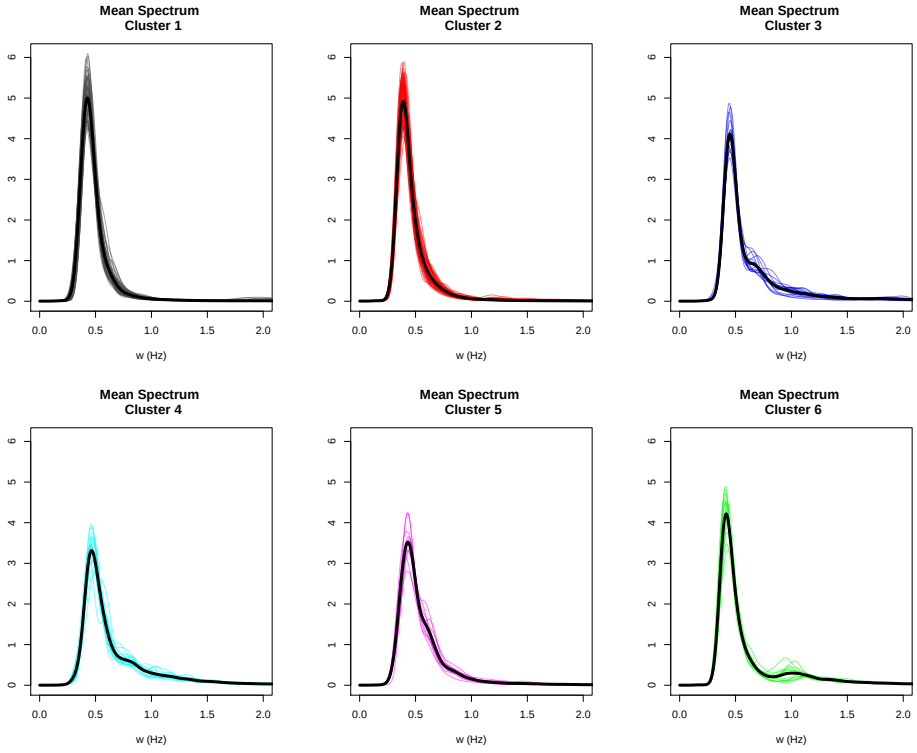
From Figure 4.6, we observe that from 0 to 27 hours, almost all segments belong to Cluster 1 (black), just a few of the members in Cluster 1 are mixed with Cluster 2 (red). This is reasonable since both spectra are unimodal and the modal frequencies are close, the one from Cluster 2 being smaller than the modal frequency of Cluster 1. Then, the members in Clusters 3 (blue), 4 (cyan) and 5 (magenta) are more mixed (in time) than the members in other clusters. These could be related with a transition between Cluster 1 and 2. Finally, Cluster 6 (green) has a bimodal spectrum, however we could not give a precise interpretation because it is close to the border and one should take a look at the following intervals.

In the case of the average link, we choose one cluster less than for the complete link case. We observe in Figure 4.7 that the clusters between 28 and 45 hours (3 and 4 in the complete link case) merge into one cluster, Cluster 3 (blue) in this case. However, some members of Cluster 1 (black) appear between Cluster 4 (cyan) and 2 (red). On the other hand, the average linkage function seems to produce clusterings that are more homogeneous in time than those obtained using the complete link, although further research in this respect is needed. Since, Cluster 5 in case 1 (magenta, when we use the complete link) and Cluster 4 in case 2 (cyan, when we use the average link) are located when H_s increases and w_p is moving, we could consider this as a transition period. So, a possible conclusion from this analysis is that there are three stable periods: 1 - 27, 28 - 45 and 52 - 89 hours, and the other intervals correspond to transition periods.

This methodology has been applied to a longer data series. Results show that the method is able to detect stable intervals, during which the distribution of the energy as a function of frequency have similar patterns, and also allows the identification of unstable or transition periods. This analysis gives statistical characteristics for the duration of stationary intervals, which may vary for different periods of the year. The complete

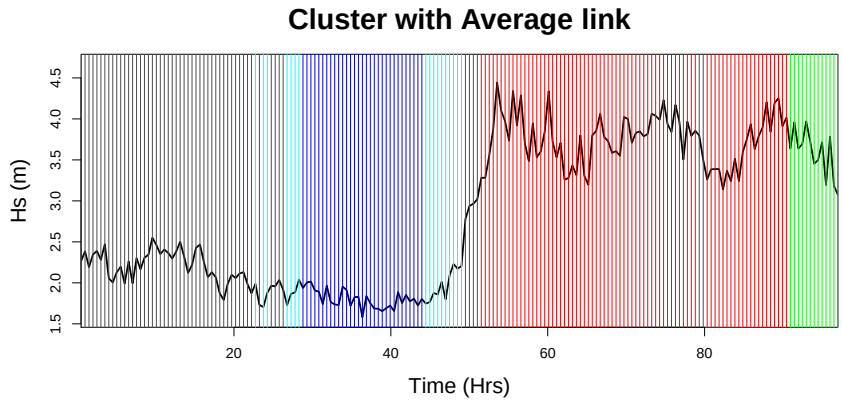


(a)

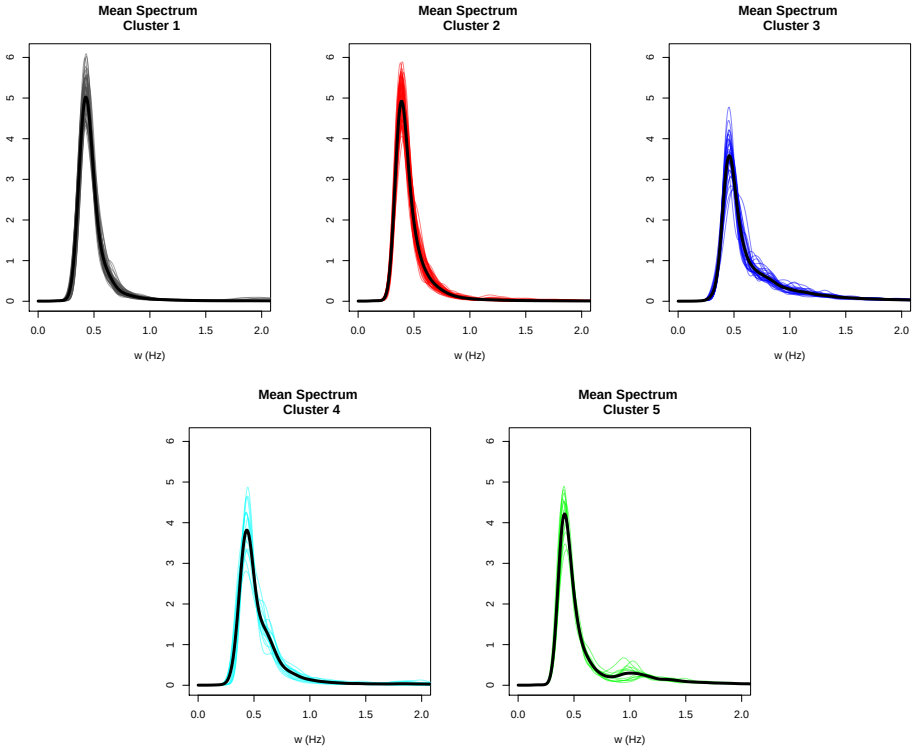


(b)

Figure 4.6: Clustering result using the complete link function and 6 clusters



(a)



(b)

Figure 4.7: Clustering result using the average link function and 5 clusters

analysis can be found in Alvarez-Esteban et al. (2016a).

4.2 Clustering of EEG data

Brain activity following stimulus presentation and during resting state are often the result of highly coordinated responses of large numbers of neurons both locally (within each region) and globally (across different brain regions). Coordinated activity of neurons can give rise to oscillations which are captured by electroencephalograms (EEG).

Spectral analysis of time series is a natural approach for studying EEG data because it identifies frequency oscillations that dominate the signal. It has many applications in neuroscience because EEG signals can be seen as a superposition of components oscillating at different frequencies. The range of frequencies that can be observed in a signal depends on the sampling frequency, usually measured in Hertz (number of cycles per second). Moreover, the convention for the different frequency bands are as follows: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-50 Hz).

The analysis of EEG data is different from the ocean waves study, since we have multichannel data and multiple trials/epochs (replicates), and the changes in brain signals can be abrupt. The HSM method will produce clusters of EEG channels according to the similarity of their spectra. The resulting clusters serve as a proxy for segmenting the brain cortical surface since the EEGs capture neuronal activity over a locally distributed region on the cortical surface.

We will analize a data set from a motor skill experiment (see Wu et al., 2014). The original study investigates how measures of cortical network function acquired at rest using dense-array EEG predict subsequent acquisition of a new motor skill. Using a partial least squares regression (PLS), they found that the coherence with the region of the left primary motor area in resting EEG was a strong predictor of motor skill acquisition. We will follow their interest in analyzing the resting state, in particular the study of the spectral profiles during rest.

Our goal is to cluster resting-state EEG signals that are spectrally synchronized, i.e., that show similar spectral profiles from subjects in this study. The participants here are healthy subjects whose EEG clustering will serve as a “standard" to which the clustering of stroke patients (with severe

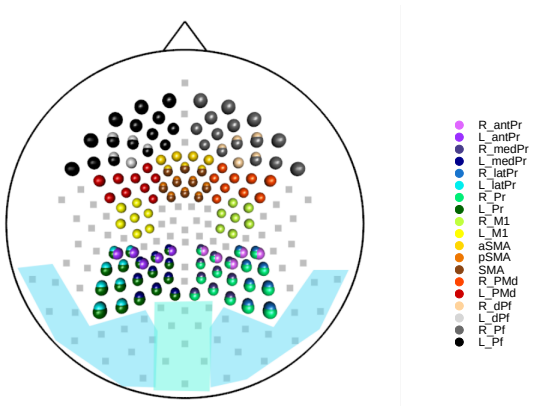


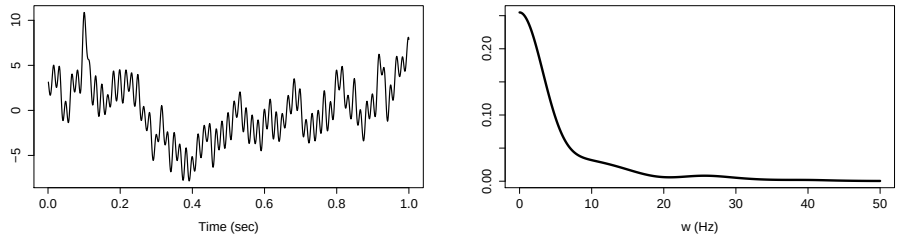
Figure 4.8: Brain regions defined in Wu et al. (2014); Left/Right Prefrontal (L_Pf, R_Pr), Left/Right Dorsolateral Prefrontal (L_dPr, R_dPr), Left/Right Pre-motor (L_PMd, R_PMd), Supplementary Motor Area (SMA), anterior SMA (aSMA), posterior SMA (pSMA), Left/Right Primary Motor Region (L_M1, R_M1), Left/Right Parietal (L_Pr, R_Pr), Left/Right Lateral Parietal (L_latPr, R_latPr), Left/Right Media Parietal (L_medPr, R_medPr), Left/Right Anterior Parietal (L_antPr, R_antPr). Gray squared channels do not belong to any of these regions; Light blue region corresponds to right and left occipital and light green region corresponds to central occipital.

motor impairment) will be compared. Some specific questions of interest are:
(1.) How many spectrally synchronized clusters are there during resting-state?
(2.) Does the number of clusters remain fixed across epochs during the entire resting-state?
(3.) Does cluster membership of the channels evolve across the entire resting-state?

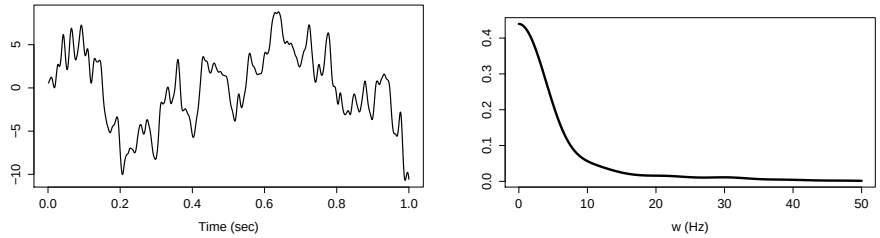
4.2.1 Data description

The EEG channels were grouped into 19 pre-defined regions in the brain as specified in Wu et al. (2014): prefrontal (left-right), dorsolateral prefrontal (left-right), pre-motor (left-right), supplementary motor area (SMA), anterior SMA, posterior SMA, primary motor region (left-right), parietal (left-right), lateral parietal (left-right), media parietal (left-right) and anterior parietal (left-right). Figure 4.8 shows the locations of these regions on the cortical surface. The number of channels for the EEG data is 256.

The data was recorded from a dense array surface using a 256-lead Hydrocel net. The complete data is formed by 17 right-handed individuals who were between 18 and 30 years of age. During the EEG-Rest period, the



(a) Channel in primary motor region



(b) Channel in prefrontal region

Figure 4.9: One second recording (1000 pts) of a brain signal and the estimated spectra.

participants were asked to hold still with the forearms resting on the anterior thigh and to direct their gaze at a fixation cross displayed on the computer monitor. Data were recorded at 1000 Hz using a high input impedance Net Amp 300 amplifier (Electrical Geodesics) and Net Station 4.5.3 software (Electrical Geodesics). Data were preprocessed. The continuous EEG signal was low-pass filtered at 100 Hz, segmented into non-overlapping 1 second epochs, and detrended. The original number of channels (256) had to be reduced to 194 because of the presence of artifacts in channels that could not be corrected (e.g. loose leads).

Smoothing the periodogram curves. To determine a reasonable value for the smoothing bandwidth, we adapted the Gamma-deviance generalised cross validation (Gamma GCV) criterion in Ombao et al. (2001) to the multi-channel setting. We applied the Gamma GCV criterion to each channel for all epochs. Trajectories of the Gamma GCV for each channel were very different because this criterion depends on the shape of the estimated spectra. There is not a common optimal bandwith for all channels. A minimum appears at

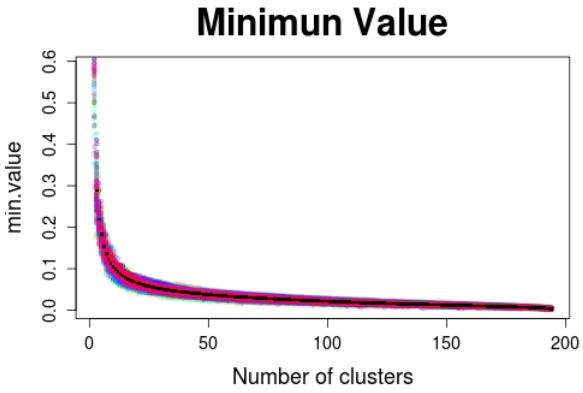


Figure 4.10: Minimum value obtained at the k-th step of the algorithm for each epoch.

$a = 80$. From the spectral estimation point of view, one could select $a = 80$ over $a = 100$. However, in our simulations, choosing the smaller bandwidth results in selecting unnecessarily too many clusters. The choice of a slightly large bandwidth, $a = 100$, gave better overall results.

4.2.2 Results using the HSM method

We analize the EEG recordings from a subject identified as "BLAK". A comparison with another subject can be found in Euán et al. (2015). The entire resting-state for each subject consisted of 160 epochs (each is a 1-second recording). Each epoch has 194 channels with $T = 1000$ time points. The HSM method (with the average version) was applied to each epoch.

To determine a reasonable number of clusters for this particular data we use the analogue of the elbow of the scree plot which in this case is the trajectory of the minimum value of the TV distance, Figure 4.10. This empirical criteria was proposed in Euán et al. (2015). To find the elbow we use the numerical derivative of the curves which is an effective visual tool for selecting the number of clusters by identifying the first value of K where the numerical derivative was below a small threshold (here, we used 0.01, based on empirical evidence from simulations). In most of the epochs, this value was equal to 9.

Now, we verify this criteria with a test based on the approximation using bootstrap for the distribution of $\hat{d}_{TV}$. In this situation, it is convenient to choose the same number of clusters in all epochs, to make them comparable,

even if, in some cases, some clusters are close to each other. The following table shows the number of epochs where the null hypothesis (9 clusters) is rejected.

α	.01	.05	.1
	0	0	2

There is not significant evidence to reject 9 clusters in any of the epochs, so, we take 9 clusters as the number of clusters for all epochs.

Even though the number of clusters remains constant across epochs, the cluster formation (i.e., location, spatial distribution, specific channel memberships) of the clusters may vary across epochs. In this EEG analysis, the total number of epochs was divided into three different phases of the resting state: early (epoch numbers 1 to 50), middle (epochs 51-110) and late (epochs 111-160).

In Figure 4.11, we show the “affinity matrix" which is the proportion of epochs when a pair of channels belong to the same cluster. The (i, j) element of the affinity matrix is the proportion of epochs such that channels i and j are clustered together – regardless of how they cluster with other channels. On the lower left corner of the affinity matrix, there are a few small red squares that represent channels that are always clustered together and completely separated from the rest.

It is evident that clustering evolved across the three phases. The affinity matrix for the early and late phases shows darker red colors which has a wider spread than that during the middle phase.

The next step in our analysis is to compare the clustering results across the different phases of resting state. Since there are 50 epochs per phase, in order to present a summary of the clustering results for each phase, we focus only on the “representative” clustering. Using the affinity matrices (in Figure 4.11), we consider the 9 clusters where the members remains most of the time in the same cluster, as the representative clustering. The procedure to get these clusters was a hierarchical cluster analysis with the complete linkage applied to the affinity matrices (considering each matrix as a similarity matrix).

Figure 4.12 shows formation of the clusters (location, spatial distribution and specific channel membership) and the shape of the corresponding spectral densities, coded in different colors, for the subject BLAK in each phase.

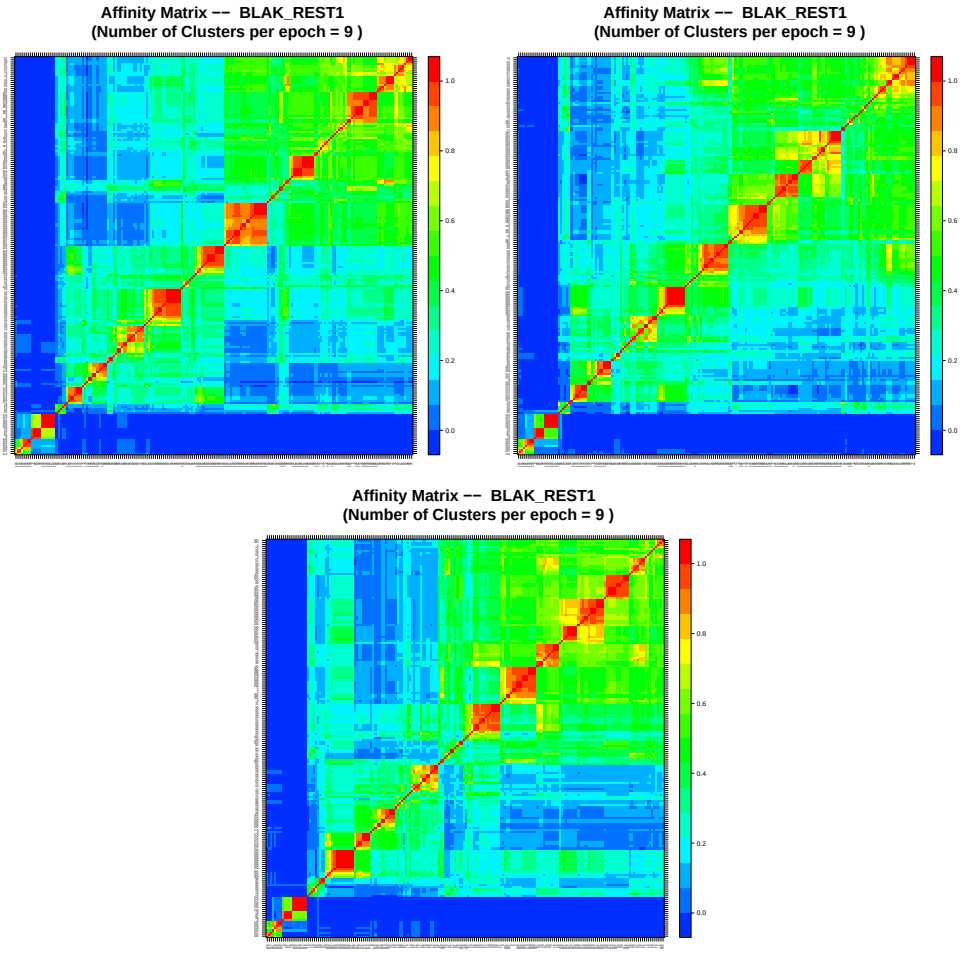


Figure 4.11: The affinity matrix: proportion of epochs where channel i and j belong to the same cluster, with 9 clusters, by segments 1-50, 51-110 and 111-160..

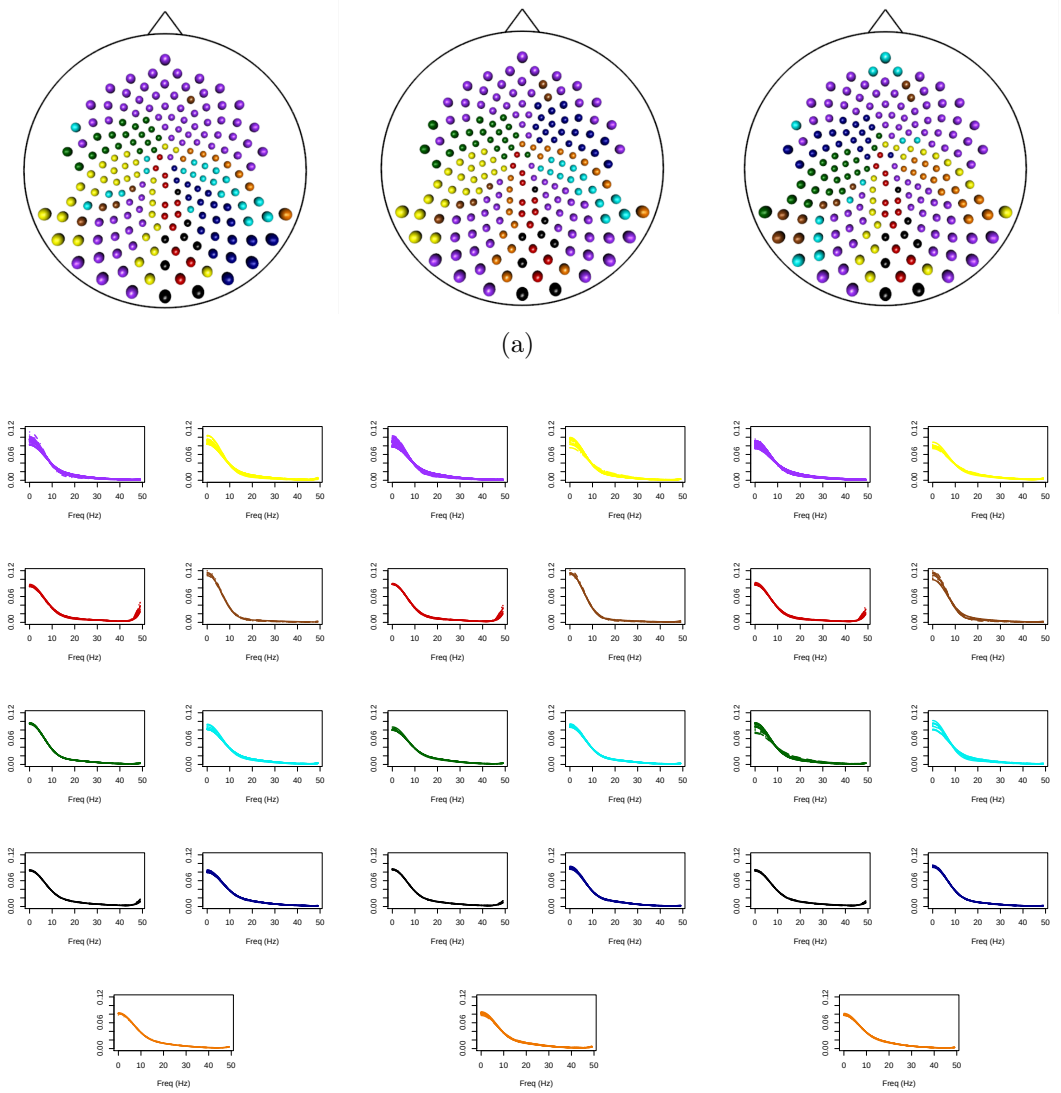


Figure 4.12: Clustering results for BLAK's resting state during different phases: early resting state (epochs 1-50), middle resting state (51-110) and late resting state (111-160). a) Distribution of clusters across the cortical surface and b) Mean spectral estimates across epochs by cluster.

Comparing the early and middle phases of resting state, we note that the formation of clusters during these phases were heavily influenced by specific bands: seven (out of the nine) clusters were dominated by the theta and alpha bands; while the formation of the remaining two clusters had also influence of the gamma band. During the late phase, the influence of the alpha band was reduced in some of the clusters but the influence of the delta and beta bands increased. The increased power in the beta band is interesting. This suggests that this subject was engaged in some cognitive task (which could not have been a response to an experimental stimulus but something that is self-induced). An study related to evidence of a relation between the beta band and attention disorder is shown in Barry et al. (2010), they report decreased levels of absolute beta and gamma power during resting state in children with attention-deficit hyperactivity disorder (ADHD), compare to healthy controls.

The formation of the clusters at the cortical surface varies across the three phases during resting state. In the early phase, channels at the left pre-motor region belong to one cluster (green) and most of the channels at prefrontal and right pre-motor region belong to another cluster (purple). However, the clustering structure at these regions changes during the middle phase where the channels in the pre-motor (which were originally clustered with the other non pre-motor channels) are assigned back with the rest of the pre-motor channels (dark blue cluster). As we transition from the middle to the late phase, channels that were assigned to the right pre-motor reverted back to the channels at the prefrontal region. These changes in cluster assignment was not entirely unexpected since many of these channels lie at the boundaries between the two anatomical regions.

Also, some channels which belong to the yellow cluster during the early phase switched to the orange cluster during the middle phase of resting-state. In this switch, the alpha and beta bands played the key roles. The late phase of the resting state shows more changes. For example, three channels located at the right occipital region switched from the yellow to the brown cluster, due to an increase of power in the alpha band and a decrease at the gamma band. Another interesting change appears on the prefontal region. There we observe three of the purple-colored channels switched to light blue cluster and a new cluster was formed. This fact lets the dark blue channels at the middle state go back to the purple ones and the underlying process that characterized this cluster (dark blue) changes completely its location.

While some channels displayed dynamic behavior across phases, there

are some clusters, such as the red and black, which showed consistent membership. The red cluster is characterized by the presence of the delta, theta bands and small activity on the gamma band while the black cluster was dominated by the theta and alpha bands.

This subject had low improvement during the task compared with the others. It is not possible to say if this has implications for the change on perceptual improvement since a causal analysis was not performed, but the presence of beta activity could produce a difference in the improvement during the task of the individuals .

The clusters produced are consistent for the most part with the anatomical-based parcellation of the cortical surface and thus cluster formation based on the spectra of the EEGs can be used to recover the spatial structure of the underlying brain process.

In addition, the HSM method has been used to analyze an epileptic seizure data. This epileptic seizure recording captures brain activity of a subject who suffered a spontaneous epileptic seizure while being connected to the EEG. The recording is digitized at 100 Hz and about 500 sec long, providing us with a time series of length T = 50000. We analyzed multichannel electroencephalograms when they exhibit "non-stationary" behavior. Our goal was to analyze the changes on the clustering of the EEG signals before, during and after the epileptic seizure. Using the HSM method we observe that only lower frequencies are mostly involved before and after the epileptic seizure (approximately 90 seconds post seizure). In contrast, immediately following seizure onset, the higher frequency bands dictated the clustering distribution of the channels. Moreover, immediately following the seizure onset but before the last subinterval, the channels were clustered similarly but the clustering was heavily influenced by the beta and gamma frequency bands. The complete analysis can be found in Ombao et al. (2016).

Conclusions

In this thesis we presented the proposal of using the total variation (TV) distance as a similarity measure in a clustering method to detect similarities in spectra.

First, we studied the theoretical properties of the TV distance between estimated spectra. We considered two asymptotic approximations of the distribution for the TV distance, a modified version for small sample size and a bootstrap procedure. The asymptotic convergence strongly depends on the election of the bandwidth, while the bootstrap procedure shows a better approximation for all bandwidth value. We established a hypothesis test which is able to detect differences in spectra and explored its power. When one uses the TV distance in a clustering method, the results are satisfactory. The rate of success of right classification is close to one and in some cases outperforms other alternatives and in other cases it is as good as other distances. The methods proposed are efficient and have shown good results in the simulation experiments.

We have used the proposed methods to analyze two different data sets, from different areas. The first analysis is related to the study of ocean waves, where the interest is to find stationary intervals. This goal was achieved using the TV distance in a hierarchical clustering algorithm, where segments in the same cluster and contiguous in time were considered as a stationary period. The second analysis is related to the study of brain signals, here we used the HSM method to detect channels of a dense EEG array that were spectrally synchronized. In both cases of study the results are good.

In general, the proposed methods have shown a good performance to detect similarities in spectra, and they can also be seen as methods to detect changes. The methodologies do not require very high computational time when one analyzes long time series or several time series. Even though we just explored two different applications, the methods are not limited to those problems and could be used in other areas.

The work performed in this project provides many possible directions for future research. They include:

- Extension of the theoretical results to a multivariate case.
- Clustering algorithms that could consider time dependency between segments or replicates,
- A more automatic method to distinguish between transitions and stable periods, in the study of oceans waves,
- Consider a windowed or weighted TV distance between different frequency bands.
- Exploring other dissimilarity measures to perform a clustering method that gives intuitive interpretations in the study of EEG data, for example "block coherence".

This problems will be studied in future research projects.

Appendices

Appendix A

R Codes

The methods developed in this project were implemented in R. We will present some of the relevant codes. The first methodology, the TV distance in a clustering algorithm, can be applied using the *hclust* package, available in the Repository (CRAN). To execute the second method, the Hierarchical Spectral Merger (HSM) method, we developed the *HSMClust* Toolbox in R.

A.1 Computing the TV distance

The *HSMClust* has one function to estimate the spectral density using the lag window estimator with a Parzen window. So, after the estimation procedure, we compute the TV distance between two normalized espectral density using the function **TVD**.

```
TVD (Total variation distance)

Description:
Computes the total variation distance between f1 and f2 with respect
to the values w using the trapezoidal rule.

Usage:      TVD(w, f1, f2)

Arguments:
w - Sorted vector of w values.

f1,f2 - Numeric vectors with the values of f1(w) and f2(w) which
```

```
are going to be compared.

*f1,f2 and w must have the same length. f1 and f2 must be
normalized functions.

spec.parzen (Smoothed periodogram using a Parzen window)

Description:
One-side estimated spectrum using a lag window estimador with a
parzen window.

Usage:
spec.parzen(x, a = 100, dt = 1, w0 = 10^(-5), wn = 1/(2 * dt), nn
= 512)

Arguments
x - Time series.

a - Bandwidth value.

dt - Sampling interval. Also, dt=1/Fs where Fs is the sampling
frequency. Default value is 1.

w0,wn -Range of frequencies of interest. By default (10^-5,Fs/2),
where Fs
is the sampling frequency.

nn - Number of evaluated frequencies in (w0,wn).

Value
A matrix of 2 columns and nn rows, where the first column corresponds
to the grid of frequencies and the second column corresponds to the
spectrum at those frequencies.
```

Examples.

```
##TVD between two normal densities
w<-seq(0,5,length=1000)
f1<-dnorm(w,2,.5)
```

```
f2<-dnorm(w,2.5,.5)
diss<-TVD(w,f1,f2)
plot(w,f1,type="l",lwd=2,col=2,main=paste("TVD =",round(diss,3)),
xlab="x",ylab="")
lines(w,f2,col=3,lwd=2)

##TVD between the normalized estimated spectra of two AR2
#processes
X1<-Sim.Ar(1000,12,1.01,100)
X2<-Sim.Ar(1000,15,1.01,100)
fest1<-spec.parzen(X1,a=300,dt=1/100)
fest2<-spec.parzen(X2,a=300,dt=1/100)
diss<-TVD(fest1[,1],fest1[,2]/var(X1),fest2[,2]/var(X2))
plot(fest1[,1],fest1[,2]/var(X1),type="l",lwd=2,col=2,
main=paste("TVD=",round(diss,3)),xlab="w (Hz)",ylab="",
ylim=c(0,max(fest1[,2]/var(X1),fest2[,2]/var(X2))))
lines(fest2[,1],fest2[,2]/var(X2),col=3,lwd=2)
```

A.2 Methods

We present the code related to the example in Chapter 3.

Example 1.

```
#################
set.seed(2786)
library(HSMClust)
normaliza<-function(f,w){
  nor<-((w[2:length(w)]-w[1:(length(w)-1)])%*%
        (f[2:length(w)]+f[1:(length(w)-1)])/2)
  return(f/nor)
}
#################

# Simulated Data
M<-1.05
eta1<-.053
```

```
eta2<-.06
Time<-1000
k<-2
nk<-3
X<-matrix(0,nrow=Time,ncol=k*nk)
for(i in seq(1,k*nk,2))X[,i]<-Sim.Ar(Time,eta1,M)
for(i in seq(1,k*nk,2)+1)X[,i]<-Sim.Ar(Time,eta2,M)


# TV distance in a clustering algorithm

# 1- Compute the dissimilarity matrix
Fest_aux<-apply(scale(X,scale=FALSE),2,spec.parzen,a=100,dt=1,nn=512)
Fest<-Fest_aux[513:1024,]
w<-Fest_aux[1:512,1]
matplot(w,Fest,type="l",lwd=3,xlim=c(0,.2),xlab="w (Hz)",ylab="",col=1,
        main="Estimated Spectra",lty=1)
FestMN<-apply(Fest,2,normaliza,w=w)
S<-matrix(0,k*nk,k*nk)
for(i in 1:(k*nk))for(j in i:(k*nk))S[i,j]<-TVD(w,FestMN[,i],FestMN[,j])
S[lower.tri(S)]<-t(S)[lower.tri(S)]

# 2- Execute the hierarquical algorithm
library(cluster)
library(dendroextras)
require(clv)
arbol<-agnes(S,diss=TRUE,method='complete',keep.diss=200)
arbol2<-agnes(S,diss=TRUE,method='average',keep.diss=200)

# 3- Results
clus<-slice(as.dendrogram(arbol),k=2)
clus2<-slice(as.dendrogram(arbol2),k=2)

# HSM Method

ClustHSM<-HSM(X)
cutk(HSM,2)
```

The HSM method is implemented in the function HSM and we get k clusters with the cutk function.

```
HSM (Hierarchical spectral merger algorithm)

Description:
Compute the hierarchical merger clustering algorithm or
the hierarchical spectral merger clustering algorithm
for a set of time series X.

Usage:
HSM(X, freq = 1, Merger = 1, par.spectrum = c(100, 1/(2 * dt), 512))

Arguments

X - Matrix of time series, the series should be located by column.

freq - Sampling Frequency. Default value is 1.

Merger - If Merger==1 (default), the algorithm will estimate the
new spectral density with the concatenated signals in order to get
a better estimation of the original spectral density. If Merger==2
the algorithm will estimate the new spectral density with the mean
spectrum using all time series in the cluster.

par.spectrum - A vector of length 3 with the parameters for
the estimation:
                par.spectrum[1]=Bandwidth value,
                par.spectrum[2]=maximun evaluated frequency,
                par.spectrum[3]= length of the grid of the
                frequencies values.


Value:
A HSM object with the following variables:
Diss.Matrix = Initial dissimilarity matrix.
min.value = trayectory of the minimum value.
Groups = list with the groupping structure at each step.
```

```
cutk (K groups from HSM)

Description:
Returns k groups from a HSM object.

Usage
cutk(Clust, kg = NA, alpha = NA)
Arguments

Clust - Output from HSM.

kg - Number of groups.

alpha - TVD value before the next clustering step.
```

Appendix B

Effect of Sampling Frequency

If a stationary process $X(t)$ with continuous spectral density, $f(\omega)$, is sampled with a sampling frequency F_s, the observed sequence is

$$X_d(t) = \sum_{m=-\infty}^{\infty} X(t)\delta(t - mdt), \tag{B.1}$$

where $dt = 1/F_s$ and $\delta(u)$ is the impulse function or Dirac delta function, which satisfies that

$$\delta(u) = 0, \text{ if } u \neq 0 \qquad \text{and} \qquad \int \delta(u)\mathrm{d}u = 1.$$

Then, the spectral density of the discrete signal can be written as a folding of the original spectral density,

$$f_d(\omega) = \sum_{m=-\infty}^{\infty} f(\omega - m/dt),$$

for $0 < \omega \leq \frac{F_s}{2}$.

Remark. Notice that we cannot observe the presence of frequencies bigger than $\frac{F_s}{2}$, since we need more than two observe time points to observed a complete period.

B.1 Discrete Fourier Transform

The periodogram is the modulus of the Discrete Fourier Transform (DFT), therefore, it is important to make some remarks on the estimation procedure

when the sampling frequency is different to one (see Mandal and Asif, 2007). Consider the Fourier transform of (B.1), using the following calculation. Let $\hat{X}_d(\omega)$ be the Fourier transform of the discrete signal, $\hat{X}(\omega)$ the Fourier transform of the continuous signal $X(t)$ and $\hat{g}(\omega)$ the Fourier transform of $\sum_{m=-\infty}^{\infty} \delta(t - mdt)$. Then,

$$\begin{aligned}
\hat{X}_d(\omega) &= \hat{X}(\omega) * \hat{g}(\omega) \\
&= \hat{X}(\omega) * \frac{2\pi}{dt} \sum_{m=-\infty}^{\infty} \delta\left(\omega - m\frac{2\pi}{dt}\right) \qquad \text{(B.2)} \\
&= \frac{2\pi}{dt} \sum_{m=-\infty}^{\infty} \hat{X}(\omega) * \delta\left(\omega - m\frac{2\pi}{dt}\right) \\
&= \frac{2\pi}{dt} \sum_{m=-\infty}^{\infty} \int \hat{X}(\omega)\delta\left(\omega - \omega * -m\frac{2\pi}{dt}\right) \mathrm{d}\omega^* \\
&= \frac{2\pi}{dt} \sum_{m=-\infty}^{\infty} \hat{X}\left(\omega - m\frac{2\pi}{dt}\right). \qquad \text{(B.3)}
\end{aligned}$$

To obtain (B.2), we have to use the Fourier transform of $\delta(t - mdt)$ which is

$$\frac{2\pi}{dt}\delta\left(\omega - m\frac{2\pi}{dt}\right).$$

To obtain (B.3) we use a property of the Dirac function,

$$\int h(t)\delta(t - \tau)\mathrm{d}t = h(\tau).$$

Finally, we get that

$$\hat{f}_d(\omega) = \frac{2\pi}{dt} \sum_{m=-\infty}^{\infty} f\left(\omega - m\frac{2\pi}{dt}\right).$$

Hence, when we use the Fourier transform of the signal (periodogram) to estimate the spectral density, we need to consider a scaling factor which is $\frac{dt}{2\pi}$, i.e., the periodogram in case of a sampling frequency should be

$$\frac{dt}{2\pi}\hat{X}_d(\omega).$$

Bibliography

Aage, C., Allan, T., Carter, D., Lindgren, G., and Olagnon, M. (1999). *Oceans from Space: A textbook for offshore engineers and naval architects*. Edition Ifremer.

Alvarez-Esteban, P. C., Euán, C., and Ortega, J. (2016a). Statistical analysis of stationary intervals for random waves. In *In Proceedings of the 26th International Offshore and Polar Engineering Conference (to appear)*.

Alvarez-Esteban, P. C., Euán, C., and Ortega, J. (2016b). Time series clustering using the total variation distance with applications in Oceanography. *Environmetrics (to appear)*.

Alvarez-Esteban, P. C., Matrán, C., del Barrio, E., and Cuesta-Albertos, J. A. (2012). Similarity of samples and trimming. *Bernoulli*, 18(2):606–634.

Barry, R., Clarke, A., Hajos, M., McCarthy, R., Selikowitz, M., and Dupuy, F. (2010). Resting-state eeg gamma activity in children with attention-deficit/hyperactivity disorder. *Clinical Neurophysiology*, 121(11):1871–1877.

Basalto, N. and De Carlo, F. (2006). *Practical fruits of econophysics: proceedings of the third nikkei econophysics symposium*, chapter Clustering financial time series, pages 252–256. Springer Tokyo, Tokyo.

Bengtsson, T. and Cavanaugh, J. E. (2008). State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. *Environmetrics*, 19(2):103–121.

Bloomfield, P. (1976). *Fourier analysis of time series: an introduction*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley.

Brillinger, D. R. (1981). *Time series: data analysis and theory*. Holden-Day, Inc., Oakland, Calif., second edition.

Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods*. Springer, New York. Reprint of the second (1991) edition.

Brodtkorb, P. A., Johannesson, P., Lindgren, G., Rychlik, I., Rydén, J., and Sjö, E. (2011). *WAFO - a matlab toolbox for analysis of random waves and loads*. Mathematical Statistics, Centre for Mathematical Sciences, Lund University.

Caiado, J., Maharaj, E. A., and D'Urso, P. (2015). *Handbook of Cluster Analysis*, chapter Time Series Clustering, pages 241–263. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

Corduas, M. (2011). Clustering streamflow time series for regional classification. *Journal of Hydrology*, 407(1–4):73 – 80.

Cuesta-Albertos, J. A. and Fraiman, R. (2007). Impartial trimmed k-means for functional data. *Computational Statistics and Data Analysis*, 51(10):4864 – 4877.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37.

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *The Annals of Statistics*, 28(6):1762–1794.

Dahlhaus, R. (2011). Locally Stationary Processes. *ArXiv e-prints*.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

Dette, H. and Hildebrandt, T. (2012). A note on testing hypotheses for stationary processes in the frequency domain. *Journal of Multivariate Analysis*, 104(1):101 – 114.

Dette, H. and Paparoditis, E. (2009). Bootstrapping frequency domain tests in multivariate time series with an application to comparing spectral densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):831–857.

Dietrich, C. R. and Newsam, G. N. (1997). Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107.

Euán, C., Ombao, H., and Ortega, J. (2015). Spectral synchronicity in brain signals. *arXiv:1507.05018v1*.

Euán, C., Ortega, J., and Alvarez-Esteban, P. C. (2014). Detecting stationary intervals for random waves using time series clustering. In *Proceedings of the 33rd. International Conference on Ocean and Arctic Engineering*, pages 1–7. ASME.

Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R. (2000). Mining the stock market: which measure is best. In *In proceedings of the 6 th ACM International Conference on Knowledge Discovery and Data Mining*, pages 487–496.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

Gnedenko, B. V. and Kolmogorov, A. N. (1968). *Limit distributions for sums of independent random variables.* Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills., Ont.

Hasselmann, K., Barnett, T., Bouws, E., Carlson, H., Cartwright, D., Enke, K., Ewing, J., Gienapp, H., Hasselmann, D., Kruseman, P., Meerburg, A., Mller, P., Olbers, D., Richter, K., Sell, W., and Walden, H. (1973). Measurements of wind-wave growth and swell decay during the joint north sea wave project (jonswap). Deutschen Hydrographischen Zeitschrift 12, Deutsches Hydrographisches Institut Hamburg.

Jentsch, C. and Pauly, M. (2012). A note on using periodogram-based distances for comparing spectral densities. *Statistics and Probability Letters*, 82(1):158–164.

Kreiss, J.-P. and Paparoditis, E. (2015). Bootstrapping locally stationary processes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 77(1):267–290.

Lachiche, N., Hommet, J., Korczak, J., and Braud, A. (2005). Neuronal clustering of brain fmri images. *Pattern Recognition and Machine Intelligence: Lecture Notes in Computer Science*, 3776:300–305.

Last, M. and Shumway, R. (2008). Detecting abrupt changes in a piecewise locally stationary time series. *Journal of Multivariate Analysis*, 99(2):191–214.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1):79–102.

Lavielle, M. and Ludeña, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli*, 6(5):845–869.

Leone, F. C., Nelson, L. S., and Nottingham, R. B. (1961). The folded normal distribution. *Technometrics*, 3(4):543–550.

Liao, T. W. (2005). Clustering of time series data – a survey. *Pattern Recognition*, 38:1857–1874.

Longuet-Higgins, M. S. (1957). The statistical analysis of a random, moving surface. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 249(966):321–387.

Maharaj, E. A. and D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Information Sciences*, 181(7):1187 – 1211.

Mandal, M. and Asif, A. (2007). *Continuous and discrete time signals and systems*. Cambridge University Press, New York, first edition.

Montero, P. and Vilar, J. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1).

Ochi, M. K. (1998). *Ocean waves: the stochastic approach*. Cambridge, U.K. ; New York : Cambridge University Press.

Ombao, H. and Bellegen, S. V. (2008). Evolutionary coherence of nonstationary signals. *IEEE Transactions Signal Process.*, 56(6):2259–2266.

Ombao, H., Schröder, A. L., Euán, C., Ting, C.-M., and Samdin, B. (2016). *Handbook of Neuroimaging Data Analysis*, chapter Advanced topics for modeling electroencephalograms (to appear), pages 567–621. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

Ombao, H., von Sachs, R., and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531.

Ombao, H. C., Raz, J. A., Strawderman, R. L., and Sachs, R. V. (2001). A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika*, 88(4):1186–1192.

Paparoditis, E. (2010). Validating stationarity assumptions in time series analysis by rolling local periodograms. *Journal of the American Statistical Association*, 105(490):839–851.

Pértega Díaz, S. and Vilar, J. A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of Classification*, 27(3):333–362.

Pierson, W. J. (1955). Wind generated gravity waves. volume 2 of *Advances in Geophysics*, pages 93 – 178. Elsevier.

Preuss, P., Vetter, M., and Dette, H. (2013). Testing semiparametric hypotheses in locally stationary processes. *Scandinavian Journal of Statistics. Theory and Applications*, 40(3):417–437.

Priestley, M. B. (1981). *Spectral analysis and time series. Vol. 1.* Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London-New York. Univariate series, Probability and Mathematical Statistics.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Rudolph, P. E. (1995). Permutation tests: A practical guide to resampling methods for testing hypotheses. *Biometrical Journal*, 37(2):150–150.

Savvides, A., Promponas, V. J., and Fokianos, K. (2008). Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition*, 41(7):2398 – 2412.

Sergides, M. and Paparoditis, E. (2008). Bootstrapping the local periodogram of locally stationary processes. *Journal of Time Series Analysis*, 29(2):264–299.

Shumway, R. H. and Stoffer, D. S. (2011). *Time series analysis and its applications. With R examples*. Springer, New York, third edition.

Torsethaugen, K. (1993). A two-peak wave spectrum model. In *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering (OMAE)*, volume II, pages 175–180.

Torsethaugen, K. and Haver, S. (2004). Simplified double peak spectral model for ocean waves. In *Proceedings of the 14th International Offshore and Polar Engineering Conference*, pages 23–28.

Wu, J., Srinivasan, R., Kaur, A., and Cramer, S. C. (2014). Resting-state cortical connectivity predicts motor skill acquision. *NeuroImage*, 91:84–90.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.