Wavelet Methods for Multivariate Nonstationary Time Series

Timothy Alexander Park, M.Phys, M.Res

Submitted for the degree of Doctor of Philosophy at

Lancaster University.

September 2014

STOR-i

excellence with impact

Abstract

This thesis proposes novel methods for the modelling of multivariate time series. The work presented falls into three parts. To begin we introduce a new approach for the modelling of multivariate non-stationary time series. The approach, which is founded on the locally stationary wavelet paradigm, models the second order structure of a multivariate time series with smoothly changing process amplitude. We also define wavelet coherence and partial coherence which quantify the direct and indirect links between components of a multivariate time series. Estimation theory is also developed for this model.

The second part of the thesis considers the application of the multivariate locally stationary wavelet framework in a classification setting. Methods for the supervised classification of time series generally aim to assign a series to one class for its entire time span. We instead consider an alternative formulation for multivariate time series where the class membership of a series is permitted to change over time. Our aim therefore changes from classifying the series as a whole to classifying the series at each time point to one of a fixed number of known classes. We also present asymptotic consistency results for this framework.

The thesis concludes by introducing a test of coherence between components of a multivariate locally stationary wavelet time series.

Acknowledgements

I would like to thank the STOR-i Centre for Doctoral Training for providing me with a pleasant working environment. In particular I would like to thank my supervisor Idris Eckley. I would also like to thank my collaborator Hernando Omabo for his advice and support.

I gratefully acknowledge the financial support of both EPSRC and Unilever Research. I would also like to thank my industrial supervisor at Unilever's Port Sunlight Laboratory, Ruediger Zillmer for giving me insight into some of the real applications of statistics.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Timothy Park

List of Papers

This thesis contains chapters which have been published or submitted for publication as follows:

Chapter 3: Park, T., Eckley, I., Ombao, H. (2014). Estimating Time-Evolving Partial Coherence Between Signals via Multivariate Locally Stationary Wavelet Processes. *IEEE Transactions on Signal Processing* 62(20):5240-5250.

Chapter 4: Park, T., Eckley, I., Ombao, H. (2014). Dynamic Classification using Multivariate Locally Stationary Wavelets. *In Submission to IEEE Transactions on Signal Processing*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The wavelet transform introduced by Daubechies (1990) has received considerable attention within the statistics community over the last twenty years. Much of their utility derives from their localised form which permits a location dependent frequency decomposition of a function, time series or image. This allows for more efficient modelling of features such as gradual changes in structure or sudden discontinuities. Consequently wavelet based methods have been applied to many different classes of problems in areas such as time series, signal processing and image processing.

Within the wavelets time series literature, one of the key developments has been the introduction of the locally stationary wavelet process by Nason et al. (2000). Their model makes use of the localisation of the wavelet basis to allow for smooth changes in the second order structure of a time series, thus removing the need to assume stationarity. Removing this, often restrictive, assumption means that the model of Nason et al. (2000) can be applied to a wider range of time series. We will describe their approach in Chapter 2 as well as reviewing some of its recent applications in the

statistics literature. Chapter 2 also covers some of the key aspects of wavelet theory as well as contrasting this with the Fourier basis.

The more recent expansion of sensor based data recording means that the problem of modelling a multivariate time series is becoming increasingly important. Such series are often long in length and characterised by evolving properties. While univariate methods can be used to model the individual components separately this does not allow for the modelling of dependencies *between* components. To this end Chapter 3 introduces the *multivariate locally stationary wavelet* model which is able to model a multivariate time series with an evolving second order structure. This model is able to not only capture the dependencies between the components of a series but is also able to distinguish between components with a direct dependence and those which are dependent only through other components.

In Chapter 4 we make use of the multivariate locally stationary wavelet model introduced in Chapter 3 and apply it to the problem of dynamic classification of time series. The problem of classifying an *entire* time series into one of a known number of classes has been well studied in the literature, here we consider a variant on this problem. Specifically we focus on the situation where the class membership of a time series is permitted to change over time. Under this dynamic framework the class membership of a time series is estimated locally rather than globally.

Another application of the multivariate locally stationary wavelet model is covered in Chapter 5. We have already stated that the multivariate locally stationary wavelet model can be used to identify dependencies between different components in a multivariate time series. In Chapter 5 we introduce a formal hypothesis test for

coherence which aims to determine if these dependencies are statistically significant. This makes it possible to easily identify which components are dependent and, by using a wavelet basis, we are also able to identify which time and frequency points are contributing to the dependence. Finally in Chapter 6 we conclude with some ideas for future research.

Chapter 2

Literature Review

2.1 Introduction

In this chapter we review some of the key concepts of time series analysis which we will build upon in later chapters. We define a **time series**, which is also referred to as a **signal**, to be a set of observations, X_t, of a process measured sequentially through time. These measurements can either be made continuously through time or at a discrete set of time points. Within this thesis we restrict ourselves to discrete time observations $X_t,\ t \in \mathbb{N}$ where $X_t \in \mathbb{R}$.

Typically such time series display some degree of serial dependence, i.e. the value of the series at time t will depend on the value of the process at previously observed time points. Examples of such processes include the well-known moving average processes of order p (often denoted MA(p)) which takes the form, $X_t = \xi_t + \sum_{i=1}^{p} \theta_i \xi_{t-i}$, where $\{\theta_i\}_{i\in\{1,\ldots,p\}}$ is the set of model parameters and $\{\xi_t\}_{t\in\mathbb{N}}$ is a set of independent and identically distributed zero-mean random innovations. Another time series

model which we can use is the autoregressive (AR) processes. Typically an AR process of order q is denoted as AR(q) and takes the form, $X_t = \sum_{i=1}^{q} \phi_i X_{t-i} + \xi_t$, where $\{\phi_i\}_{i\in\{1,\ldots,q\}}$ is the set of model parameters. Many excellent texts have been written on the subject of time series analysis. We therefore refer interested readers to Priestley (1981a); Shumway and Stoffer (2000); Chatfield (2003) and Brockwell and Davis (2009) for a comprehensive treatment of these long-established time series models.

One of the key concepts which has underpinned much of the previous work on time series modelling is that of **stationarity**. A time series is said to be *strictly stationary* if the joint distribution of a set of time series observations, $X_1, \ldots, X_n$ is identical to the joint distribution of the observations, $X_{1+\tau}, \ldots, X_{n+\tau}$, for some value of $\tau \in \mathbb{Z}$. An alternative, less restrictive assumption is that a time series is *second order stationarity*. A time series is said to be second order stationary if it has a constant mean and the covariance between observations only depends upon the lag between them, so that,

$$E[X_t] = \mu, \quad \text{and}, \quad \text{cov}(X_t, X_{t+\tau}) = \kappa_\tau.$$

In essence stationarity requires the key statistical properties of a time series to remain constant over time.

Note that in the remainder of this thesis, we will use the term stationary time series to mean a second order stationary time series. The remainder of this chapter proceeds as follows: Section 2.2 describes the Fourier representation of a stationary time series both in the univariate and multivariate settings. Section 2.3 describes some adaptations of the stationary Fourier basis which can be used to represent

nonstationary time series. Section 2.4 introduces the wavelet basis and gives details of different forms of wavelet transform. Section 2.5 concludes this chapter by describing the locally stationary wavelet model for nonstationary time series. Sections 2.5 also includes some recent applications of the LSW model.

2.2 Fourier Representation of a Stationary Time Series

The Fourier basis is a long established basis which can be used to construct a time series representation. In this section we give a brief overview of some widely used Fourier representations for univariate time series which are second order stationary, a more thorough description can be found in Priestley (1981a), Bloomfield (2000) or Shumway and Stoffer (2000).

The Fourier basis is essentially a combination of sine and cosine functions. Often these are combined into a single complex exponential,

$$\exp(i\omega t) = \cos(\omega t) + i\sin(\omega t).$$

Clearly this is an oscillatory function, the frequency of oscillation can be controlled by varying the parameter, ω. The Fourier basis is therefore ideal for representing series which exhibit some form of oscillation or periodicity.

2.2.1 Stationary Fourier Representation

Given the definition of the Fourier basis it is simple to construct a representation for a time series. Let X_t be a second order stationary time series with zero mean. Following the notation of Dahlhaus (1997) a Fourier representation of X_t is,

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t)\, \mathrm{d}\xi(\omega), \tag{2.1}$$

where $\mathrm{d}\xi(\omega)$ is a stochastic process with the properties that $E\left[\,\mathrm{d}\xi(\omega)\right] = 0$ and $E\left[\,\mathrm{d}\xi(\omega)\overline{\mathrm{d}\xi(\lambda)}\right] = \delta_{\omega\lambda}$. The complex valued function $A(\omega)$ is know as the amplitude or *transfer function*. For X_t to be real valued then the transfer function must have the property that $A(\omega) = \overline{A(-\omega)}$.

The second order structure of X_t can be uniquely defined in terms of its *spectrum*, $f_X(\omega)$. Let $\kappa_{X,s} = \mathrm{cov}\,(X_t, X_{t+s})$ be the covariance function for the process X_t. The spectrum at frequency ω is defined as,

$$f_X(\omega) = T \sum_{s=-\infty}^{\infty} \kappa_{X,s} \exp\left(-i\omega s T\right), \tag{2.2}$$

Inverting this relationship we see that the covariance can be expressed in terms of the spectrum as follows,

$$\kappa_{X,s} = \frac{1}{T} \int f_X(\omega) \exp\left(i\omega s T\right) \mathrm{d}\omega. \tag{2.3}$$

Expressing the covariance in this way demonstrates how the spectrum provides a frequency based decomposition of the second order structure of the time series. If we consider the variance of the series it is also possible to show that the spectrum can

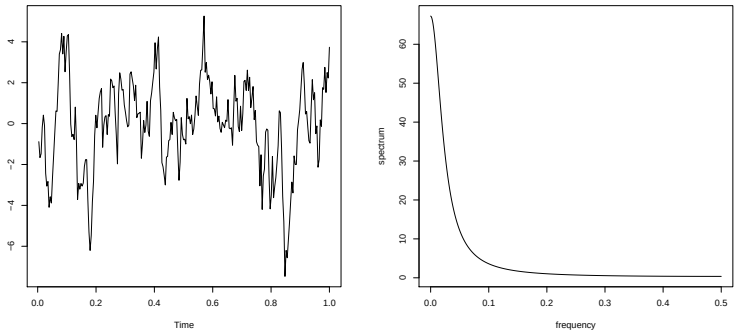
be expressed in terms of the transfer function,

$$\begin{aligned}\mathrm{Var}\{X_t\} =& E\left[X_t^2\right], \\ &= \int_{-\infty}^{\infty}\int_{\infty}^{\infty} A(\omega)\overline{A(\omega')}\exp\left[i(\omega'-\omega)t\right]E\left[\,\mathrm{d}\xi(\omega)\,\mathrm{d}\xi(\omega')\right], \\ &= \int_{-\infty}^{\infty}|A(\omega)|^2\,\mathrm{d}\xi(\omega) = \int_{-\infty}^{\infty} f_X(\omega)\,\mathrm{d}\xi(\omega).\end{aligned}$$

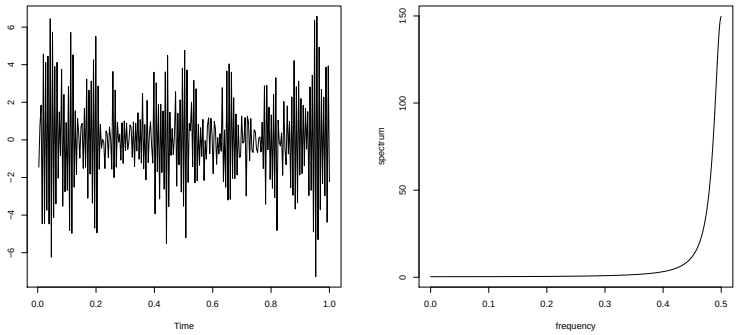
Therefore $f_X(\omega) = |A(\omega)|^2$ is an alternative definition of the spectrum which demonstrates the link between the transfer function and the second order structure.

We demonstrate the link between the spectrum and covariance structure by considering two processes X_t and Y_t. Both are realisations of an AR(1) processes with parameters of 0.9 and -0.9 respectively, i.e. $X_t = \alpha X_{t-1} + \xi_t$ where $\alpha = 0.9$ or -0.9. These series are show in Figures 2.1(a) and 2.1(c). For X_t the AR parameter of 0.9 makes it likely that consecutive time points will be close together causing the value of the series to change slowly or in other words the series is characterised by low frequencies. This is reflected in the spectrum, shown in Figure 2.1(b), which has high values in the low frequency range and near zero values for all other frequencies. Conversely for Y_t each value is likely to have the opposite sign to the previous value and so the series changes quickly and will be characterised by high frequencies. The spectrum of Y_t, Figure 2.1(d), is therefore nonzero only for high frequencies.

Spectral Estimation In order to perform inference on a series we must be able to estimate its spectrum. For a time series X_t observed at time points $t \in \{0, \ldots, T-1\}$, the first step is to take a Fourier transform of the series to obtain the set of Fourier



(a) X_t an AR(1) Process, $\alpha = 0.9$ (b) Spectrum of X_t



(c) Y_t an AR(1) Process, $\alpha = -0.9$ (d) Spectrum of Y_t

Figure 2.1: Examples of the Spectra of two different processes

coefficients, $\{b_{\omega_k}\}$. These coefficients are given by,

$$b_{\omega_k} = \frac{1}{T}\sum_{t=1}^{T} X_t \exp\left(-i\omega_k t\right) \tag{2.4}$$

where $\omega_k = \frac{k}{T}$, for $k = 0, 1, \ldots, T-1$. The Fourier coefficients are then used to compute the *periodogram* of the series, which is defined as $I(\omega_k) = |b_{\omega_k}|^2$. It can be shown that for a Gaussian time series the periodogram follows a scaled chi-squared

distribution such that,

$$\frac{I(\omega_k)}{f(\omega_k)} \sim \begin{cases} \chi_1^2 & \text{if} \quad k = 0 \\ \frac{1}{2}\chi_2^2 & \text{if} \quad k \neq 0 \end{cases} \tag{2.5}$$

See Priestley (Chapter 6, 1981) for details. Using this distributional property the periodogram can be seen to be an unbiased but inconsistent estimator of the spectrum. In order to overcome the problem of inconsistency the periodogram is generally smoothed over frequency, for example using kernel smoothing. The kernel smoothed periodogram is defined as,

$$\tilde{I}(\omega_k) = \sum_{j=-h}^{h} w(h) I(\omega_{k+j}) \tag{2.6}$$

Where $H = 2h + 1$ is the width of the smoothing kernel and $w(h)$ is the kernel function. There has been much work on selecting the optimum window size, see for example the methods of Lee (1997) or Ombao et al. (2001). A wider window lowers the variance of the estimator, however this is achieved at a cost of introducing bias.

2.2.2 Multivariate Stationary Fourier Representation

We next turn to consider an extension of the Fourier stationary time series setting for multivariate time series. While individual components of a multivariate time series can be represented separately using the univariate representation, this does not take into account any dependencies *between* components. The remainder of this section will describe two quantities which can be used to measure such dependencies: *coherence* and *partial coherence*.

Coherence: When considering a multivariate second order stationary time series it is possible that different components of the series will have some cross-dependence. One possible measure of linear dependence between components is the correlation. For a bivariate series $\boldsymbol{X}_t = [X_t^1, X_t^2]'$, whose two components have standard deviations, σ_1 and σ_2 respectively, the correlation between the two components at a lag $s \in \{1, \ldots, T-s\}$ is given by,

$$r_{12}(s) = \frac{\text{cov}\left(X_t^1, X_{t+s}^2\right)}{\sigma_1 \sigma_2}. \tag{2.7}$$

Correlation is simply the covariance at a lag of s normalised by the product of the standard deviations. The issue with time domain measures such as correlation is that they do not reveal if the relationship between the components has a frequency dependence. Identifying frequency dependence is important in many applications such as electroencephalogram (EEG) analysis where the frequency at which components are related often reveals much about the physical process. We therefore turn our attention to *coherence* as a frequency specific measure of the relationship between two components.

Before we can define the coherence between two components we must first consider how to represent the relationship between two components in the Fourier domain. Let $\kappa_{12,s} = \text{cov}\left(X_t^{(1)}, X_{t+s}^{(2)}\right)$ be the cross-covariance function between the components $X_t^{(1)}$ and $X_t^{(2)}$ at lag s. Then the *cross-spectrum*, $f_{12}(\omega)$, between two components at frequency ω is simply,

$$f_{12}(\omega) = T \sum_{s=-\infty}^{\infty} \kappa_{12,s} \exp\left(-i\omega s T\right).$$

The cross-spectrum can also be defined in terms of the transfer functions of the two series,

$$f_{12}(\omega) = A_1(\omega)\overline{A_2(\omega)}.$$

From this definition it is clear that the cross spectrum is complex and $f_{12}(\omega) = \overline{f_{21}(\omega)}$.

Definition 2.1 *Let $f_1(\omega)$ and $f_2(\omega)$ be the spectra for the two components of the bivariate time series $\boldsymbol{X}_t$ for the range of observable frequencies $\omega \in [-1/2, 1/2]$. Also let $f_{12}(\omega)$ be the cross spectrum between the two components. The coherence between the components, $\rho_{12}(\omega)$, is then defined as,*

$$\rho_{12}(\omega) = \frac{|f_{12}(\omega)|}{[f_1(\omega)f_2(\omega)]^{\frac{1}{2}}}, \qquad \omega \in [-1/2, 1/2]. \tag{2.8}$$

There are clear similarities between this definition of the coherence and the definition of correlation in equation (2.7). Note that coherence takes a value in the interval $[0, 1]$. A value of close to 1 indicates that there is a strong linear relationship between the structures of the two series at that particular frequency and may indicate a dependence between the series at that frequency. A value of close to zero indicates that at that frequency the components are independent.

Partial Coherence: When a multivariate time series consists of more than two components, simple pairwise coherence is not the only quantity we can consider. Consider by way of example the case of three components, $X_t^{(1)}$, $X_t^{(2)}$ and $X_t^{(3)}$. The coherence may indicate dependencies between all pairs of components. There are two possible explanations for this: either (a) there is some direct relationship between all three components or (b) two of the components are related only through their

relationship with the third. For example the coherence between $X_t^{(1)}$ and $X_t^{(2)}$ may be due to a direct dependence between them *or* it may be due to them both having a direct dependence with $X_t^{(3)}$. Coherence can not make this distinction, however *partial coherence* can. Partial coherence is a measures of the coherence between a pair of components after any linear relationships with all other observed components have been removed. Following Koopmans (1975) we define the partial coherence in terms of the partial coherency, $\gamma_{12.3}(\omega)$. Briefly the partial coherence is defined as follows,

Definition 2.2 *Let $X_t^{(1)}$, $X_t^{(2)}$ and $X_t^{(3)}$ be the three components of a tri-variate time series. Also let $f_p(\omega)$ be the Fourier spectrum for the p-th component of the series and let $f_{pq}(\omega)$ be the cross spectrum between the p-th and q-th components for $\omega \in [-1/2, 1/2]$. Finally let $\rho_{pq}(\omega)$ be the Fourier coherence between the p-th and q-th components of the time series. The partial coherency between components 1 and 2 is then defined as,*

$$\gamma_{12.3}(\omega) = \frac{\gamma_{12}(\omega) - \gamma_{13}(\omega)\overline{\gamma_{32}(\omega)}}{[(1-\rho_{13}^2(\omega))(1-\rho_{23}^2(\omega)]^{\frac{1}{2}}}. \tag{2.9}$$

Where and $\gamma_{12}(\omega)$ is the coherency which is defined as,

$$\gamma_{12}(\omega) = \frac{f_{12}(\omega)}{[f_1(\omega)f_2(\omega)]^{\frac{1}{2}}}. \tag{2.10}$$

The partial coherence is then the modulus of the partial coherency so that,

$$\rho_{12.3}(\omega) = |\gamma_{12.3}(\omega)| \tag{2.11}$$

Partial coherence can also be extended to cases where there are more than three series. Fore example $X_t^{(1)}$ and $X_t^{(2)}$ may have direct linear relationships with $X_t^{(3)}$, $X_t^{(4)}$ and

so on, but not with each other. Fourier partial coherence is covered in more detail by Koopmans (1975) and Priestley (1981b).

2.3 Fourier Representations for Nonstationary Time Series

Whilst the Fourier basis has been used to derive a spectral approach for stationary time series, in recent years researchers have sought to adapt the basis to permit modelling of time series whose second-order structure is evolving over time. In this section we will describe two such representations.

2.3.1 The Locally Stationary Fourier Representation

The first nonstationary Fourier representation was proposed by Priestley (1965) but we will describe the representation proposed by Dahlhaus (1997) as it includes a full asymptotic theory. The Dahlhaus representation is very similar to the stationary representation however the transfer function, $A(\omega)$ is replaced by the time varying transfer function, $A^0_{t,T}(\omega)$. The nonstationary series $X_{t,T}$ is thus represented as follows,

$$X_{t,T} = \int_{-\pi}^{\pi} A^0_{t,T}(\omega) \exp(i\omega t)\, \mathrm{d}\xi(\omega). \tag{2.12}$$

Although stationarity over the whole series is no longer necessary in this setting, this representation does assume that the series is *locally stationary*. What this means in practice is that when viewed over a sufficiently short time window the series can be assumed to be stationary. To achieve this some smoothness conditions must be

placed on $A^0_{t,T}$ to control its behaviour. The time varying spectrum for this process is defined in a similar way to the stationary case such that, $f_X(\omega) = \left|A^0_{t,T}(\omega)\right|^2$. As such the spectrum provides a time-frequency decomposition of the series as opposed to simply a frequency decomposition.

When examining the asymptotics of such a process Dahlhaus (1997) noted that the usual concept of increasing T corresponding to observing the process for a longer time was inadequate for nonstationary series. He therefore introduced the concept of *rescaled time* whereby the time series is always observed on the interval $u = t/T \in [0, 1]$. Under this framework increasing the number of observations corresponds to observing the same time span with increasing resolution.

To estimate the time varying spectrum of a process such as this we must take a Fourier transform of the time series which is localised in time. The Fourier transform described in equation (2.4) puts equal weight on all time points. A localised Fourier transform uses a taper function to put greater weight on time points closest to the time of interest. The localised Fourier coefficient calculated using taper function $\Psi(u)$ at rescaled time point u and frequency ω_k is,

$$d_{u,\omega_k} = \frac{1}{T}\sum_{t=1}^{T} X_t \Psi(t - uT)\exp\left(-i\omega_k t\right). \tag{2.13}$$

These coefficients can then be used to estimate the spectrum in a similar way to the stationary setting.

2.3.2 SLEX

Building on the ideas of a nonstationary Fourier model Ombao et al. (2002) introduced the Smooth Localised Exponential (SLEX) basis. This model aims to segment a series into stationary blocks whilst also allowing for neighbouring blocks to overlap. Ombao et al. (2002) achieve this using a projection operator which preserves orthonormality whilst allowing for smooth time localisation.

A SLEX basis function, $\psi_{S,\omega}(t)$, is localised to have support for discrete time block $S = \{\alpha_0 - \epsilon + 1, \ldots, \alpha_1 - \epsilon\}$ where ϵ is the size of overlap between neighbouring time blocks. The general form of this basis vector is,

$$\psi_{S,\omega}(t) = \Psi_{S,+}(t) \exp\left(i2\pi\omega\frac{t}{|S|}\right) + \Psi_{S,-}(t) \exp\left(-i2\pi\omega\frac{t}{|S|}\right), \tag{2.14}$$

where $\omega \in [-1/2, 1/2]$ is the oscillating frequency. This is equivalent to applying *two* tapers, one, $\Psi_{S,+}(t)$, to the positive frequencies and the second, $\Psi_{S,-}(t)$, to the negative frequency. These tapers have the general form,

$$\Psi_{S,+}(t) = r^2\left(\frac{t-\alpha_0}{\epsilon}\right) r^2\left(\frac{\alpha_1 - t}{\epsilon}\right),$$
$$\Psi_{S,-}(t) = r\left(\frac{t-\alpha_0}{\epsilon}\right) r\left(\frac{\alpha_0 - t}{\epsilon}\right) - r\left(\frac{t-\alpha_1}{\epsilon}\right) r\left(\frac{\alpha_1 - t}{\epsilon}\right).$$

The function $r(.)$ is known as the rising cut-off function, choices of this function can be found in Wickerhauser (1994).

The aim of the SLEX basis is to represent the nonstationarity of a series by segmenting it into stationary blocks. A segmentation which covers all time points with neighbouring blocks overlapping by ϵ is referred to as a basis. The set of all possible bases is referred to as the SLEX library. The library is divided into $J + 1$

different levels labelled $j \in \{0, 1, 2, \ldots, J\}$ and each level is divided up into 2^j blocks containing $M_j = T/2^j$ points. The block at level j and position $b \in \{0, 1, \ldots, 2^j - 1\}$ is labelled as $S(j, b)$. A SLEX library containing three levels is shown in Figure 2.2. This library contains five possible bases, one of which is shown by the shaded blocks.

To represent a series it is necessary to select a suitable basis from the library. Ombao et al. (2005) achieve this by assigning a complexity penalised cost to each block and then using the best basis algorithm (BBA) of Coifman and Wickerhauser (1992) to select the basis with the lowest total cost.

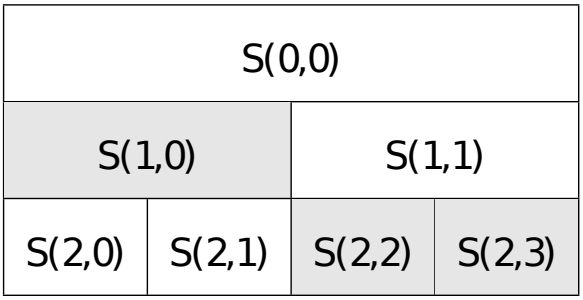


Figure 2.2: Example of a SLEX library with $J = 2$. Shaded blocks show one possible basis choice.

2.3.3 Criticisms of the Nonstationary Fourier Representation

One issue with the approach of Dahlhaus (1997) is the choice of taper function in equation (2.12). Both the span and shape of this taper must be chosen to accurately reflect the observed series. Data adaptive methods for choosing this taper exits, however these can be computationally slow. The SLEX basis overcomes this issue by including a computationally fast basis selection however this is only possible if the series is segmented dyadically. This can be very restrictive especially if the length of

the series is short. In the next section we will introduce the wavelet basis which does not need to be adapted to represent nonstationary series and so does not suffer from these problems.

2.4 Wavelet Methods

So far we have introduced the decomposition of a stationary time series in terms of its different frequency components using a Fourier basis. In this section we move to representations of nonstationary time series and introduce time-scale decompositions using a wavelet basis. We start by describing the general concept of a multiresolution analysis before introducing the wavelet basis function. We then proceed to describe some different forms of wavelet transform.

2.4.1 Multiresolution Analysis

Before we describe the properties of a wavelet basis we first describe the more general concept of *multiresolution analysis* (MRA) of a function first introduced by Meyer (1986) and Mallat (1989b). Simplistically, one might view MRA as the ability to zoom in or out in order to view a function with varying levels of detail. More precisely an MRA is a collection of closed subspaces V_n for $n \in \mathbb{Z}$ in $\mathbb{L}_2(\mathbb{R})$. These subspaces are nested such that,

$$\ldots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \ldots$$

The subspaces are constructed such that they have a trivial intersect and a dense union,

$$\cap_{j\in\mathbb{Z}} V_j = \{\mathbf{0}\}, \quad \cup_{j\in\mathbb{Z}} V_j = \mathbb{L}_2(\mathbb{R}).$$

The hierarchical structure of the subspaces means that if we define a function on one of the subspaces then we can use a dilation operation to transform the function such that it is contained within a different subspace. I.e.,

$$f(x) \in V_l \Leftrightarrow f(2^j x) \in V_{l-j} \quad \forall j \in \mathbb{Z}. \tag{2.15}$$

If we instead use a translation operator then the function remains in the same subspace,

$$f(x) \in V_j \Leftrightarrow f(x-k) \in V_j \quad \forall k \in \mathbb{Z}. \tag{2.16}$$

The different subspaces therefore allow us to view the same function, f, with different levels of dilation.

Another property of the subspaces is that there exists a scaling function $\phi \in V_0$, the integer transforms of which will form an orthonormal basis of V_0. In other words any function, $f(x) \in V_0$ can be represented as a linear combination of the integer transforms of ϕ,

$$f(x) = \sum_k c_k \phi(x-k),$$

for some set of coefficients $\{c_k\}$. Furthermore if we use the conditions in equations (2.15) and (2.16) then it is easy to show that $\{\phi(2^{-j}x - k) : k \in \mathbb{Z}\}$ is an orthonormal basis of V_j. It is also possible to show that since $V_0 \subset V_{-1}$ then we can express ϕ as

a linear combination of the functions $\phi(2x-k)$,

$$\phi = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\phi(2x-k),$$

where $\{h_k\}$ is a set of coefficients which are specific to the function ϕ.

2.4.2 Wavelet Basis Functions

Next we outline the main aspects of a wavelet basis and certain properties which make it a suitable choice for representing a nonstationary series. More details of this can be found in Vidakovic (1999) or Nason (2008).

The starting point for any wavelet basis is the *mother wavelet*, ψ. This function can be used to derive an orthonormal wavelet basis. Unlike the Fourier basis which is built using sinusoidal functions there are many different mother wavelets which can be used. In this thesis we will focus on wavelets from two families defined by Daubechies (1988). These families are referred to as "extremal phase" and "least-asymmetric" wavelets. More details of these families can be found in Daubechies (1992) or Vidakovic (1999). Within these families the different wavelet functions are characterised by the number of vanishing moments, $N \in \mathbb{N}$. A wavelet function with N vanishing moments must satisfy the following properties;

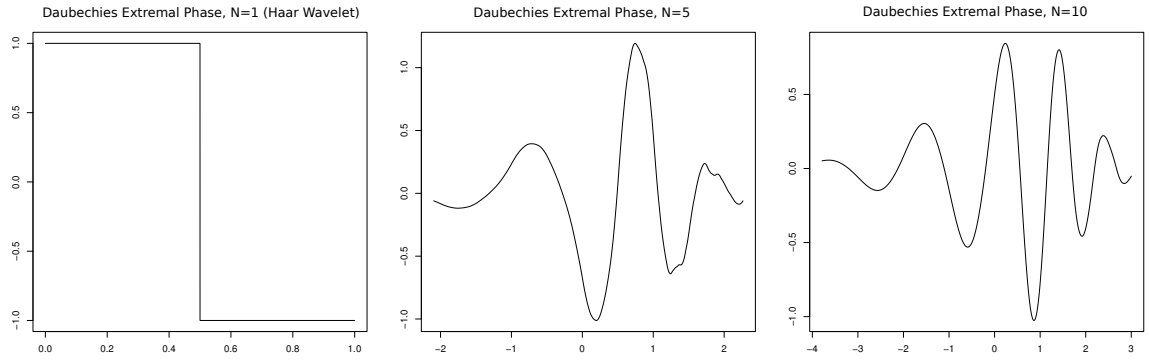
1. $\psi(x) \in L^\infty(\mathbb{R})$. Additionally if $N > 1$ then $\frac{\mathrm{d}^n}{\mathrm{d}x^n}\psi(x) \in L^\infty(\mathbb{R})$ for all $n \leq N$.

2. $\psi(x)$ and its derivatives up to order N must vanish rapidly as $x \to \pm\infty$.

3. For all $k \in \{0, \ldots, N\}$,

$$\int_{-\infty}^{\infty} x^k \psi(x)\,\mathrm{d}x = 0.$$

The second property ensures that the wavelet has compact support. This is in contrast to the standard Fourier exponential which does not vanish asymptotically. As we will demonstrate in Section 2.5 compact support is a useful property when representing nonstationary series. Some examples of Daubechies wavelets with different vanishing moments are shown in Figure 2.3.



(a) (b) (c)

Figure 2.3: Some examples of Daubechies Wavelets

An orthonormal basis can be derived from the mother wavelet using dilation and translation operators. These are represented by the coefficients j and k respectively. The set of wavelet basis functions is therefore labelled as $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$ where,

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k). \tag{2.17}$$

Looking at equation (2.17) we see that dilation can be considered as a stretching of the basis function. The parameter j is generally referred to as the level or scale. A higher value of j increases the support length of the function and gives it a longer oscillation period. In this way the dilation coefficient is similar to the frequency, ω, of the Fourier basis with higher values of j roughly corresponding to lower frequencies. In the wavelet world lower levels may also be referred to as finer scales and higher

levels as coarser scales. The translation operation changes the position of the wavelet function. This allows the basis to be localised in time as well as frequency. Using this basis we can represent a zero mean function $f(x) \in L^2(\mathbb{R})$ as a linear combination of these basis function,

$$f(x) = \sum_{j=1}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(x), \tag{2.18}$$

where $\{d_{j,k}\}$ is the set of wavelet detail coefficients. In the next section we briefly introduce various approaches for calculating these coefficients for discrete time series.

2.4.3 Discrete Wavelet Transforms

Discrete Wavelet Transform: The first type of wavelet transform we describe is the standard Discrete Wavelet Transform (DWT), proposed by Mallat (1989a). Let X_t be a discrete time series for time points $t \in \{0, \ldots, T-1\}$ where $T = 2^J$ for some $J \in \mathbb{N}$. For a series of this length the coarsest level which can be computed is level $j = J$. When taking the DWT of this series we calculate two sets of coefficients the detail coefficients, $\{d_{j,k}\}$, and the smooth coefficients, $\{c_{j,k}\}$. The finest level of the smooth coefficients can be computed directly from the series with coarser levels being computed recursively.

The formula for calculating the smooth coefficients is given as follows,

$$c_{1,k} = \sum_n h_{n-2k} X_k,$$

$$c_{j+1,k} = \sum_n h_{n-2k} c_{j,n} \quad \text{for } j \in \{1, \ldots, J\}.$$

The set $\{h_k\}$ are low pass filter coefficients which are specific to the mother wavelet

used for the transform. For level $j \in \{1, J\}$ the DWT calculates coefficients for locations $k \in \{0, \ldots, 2^{J-j} - 1\}$. The number of coefficients consequently decreases for coarser levels. A transform with this property is known as a *decimated* transform.

Similarly the detail coefficients are calculated as follows,

$$\begin{aligned} d_{1,k} &= \sum_n g_{n-2k} X_k, \\ d_{j+1,k} &= \sum_n g_{n-2k} c_{j,n} \quad \text{for } j \in \{1, \ldots, J\}, \end{aligned} \tag{2.19}$$

where $\{g_k\}$ are the high pass filter coefficients. These are again specific to the wavelet function being used and can be calculated from the low pass coefficients as,

$$g_k = (-1)^k h_{1-k}. \tag{2.20}$$

A pictorial representation of the algorithm can be seen in Figure 2.4

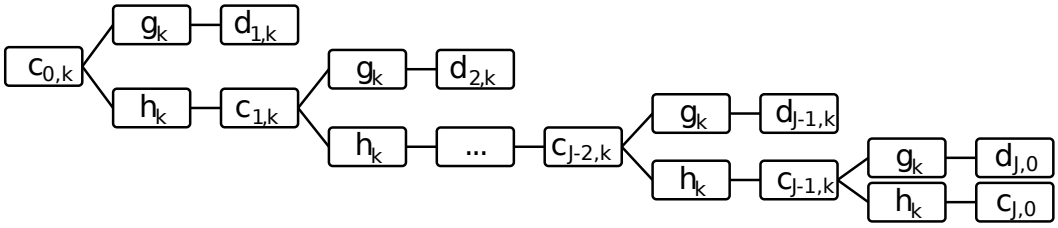


Figure 2.4: The procedure for computing coefficients from the original series

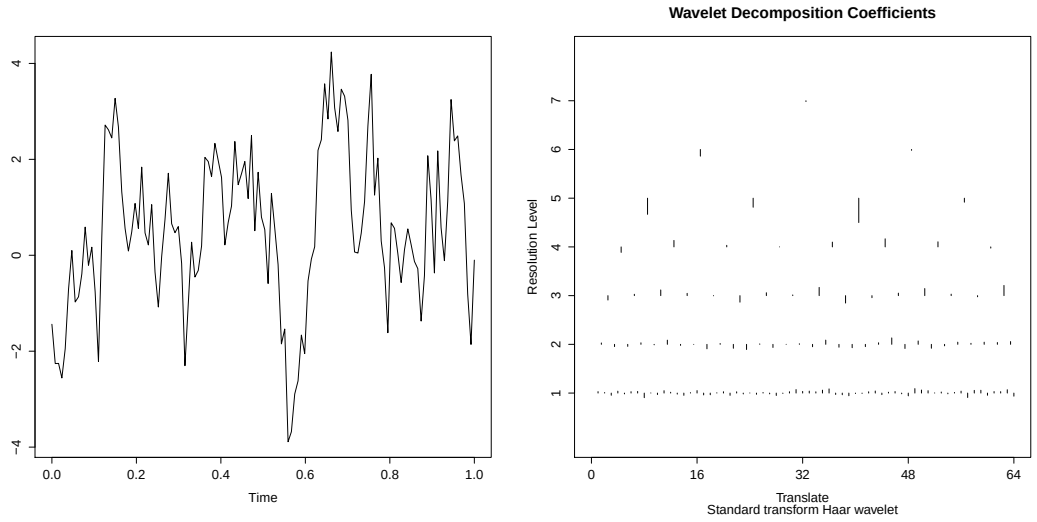
The DWT is an orthogonal transform. Consequently the inverse DWT can be computed using the following formula,

$$X_t = \sum_k h_{t-2k} c_{1,k} + \sum_k g_{n-2k} d_{1,k},$$

where,

$$c_{j,n} = \sum_k h_{n-2k} c_{j+1,k} + \sum_k g_{n-2k} d_{j+1,k} \quad j \in \{1, J-1\}.$$

As this formula shows finer scale smooth coefficients can be calculated using coarser scale smooth and detail coefficients. Consequently if the coefficients of a series are computed up to level $j_0 \leq J$ then the series can be recovered exactly using only the smooth coefficients for scale j_0 and the detail coefficients for the scales $j \in \{1, \ldots, j_0\}$. If the transform is computed up to level J then the only smooth coefficient which is needed for the inversion is $c_{J,0}$. An example of the detail coefficients of the DWT of a time series is shown in Figure 2.5.



(a) Time Series of length $T = 128$ (b) DWT Detail Coefficients.

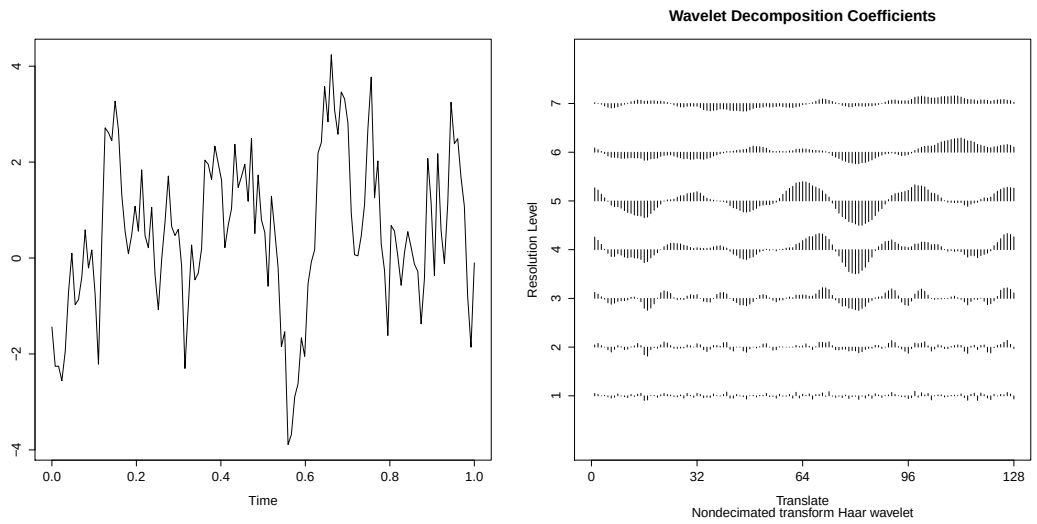
Figure 2.5: Discrete Wavelet Transform of a time series.

The plot in Figure 2.5(b) shows the detail coefficients from a DWT of the series in Figure 2.5(a). The y-axis denotes the wavelet levels from the finest level, $j = 1$, to the coarsest, $j = 7$. The x-axis denotes the location (or sequence order). The lengths of the vertical lines corresponds to the size of the detail coefficient at that

scale and location. It is interesting to see how the features of the series show up in the coefficients. For example the dip in the series just before time point $u = 0.6$ corresponds to a relatively large coefficient at level $j = 5$ in Figure 2.5(b).

Non Decimated Wavelet Transform: In the DWT described above the number of coefficients in each scale decreases as j increases. This is because at each scale the wavelet coefficients are only calculated for half of the possible locations. It is only by convention that in equation (2.19) we select the even sequence of locations, $2k$, as opposed to the odd sequence, $2k - 1$. Clearly this raises questions about what extra information might be gained by considering both the odd and even locations. The non-decimated wavelet transform, NDWT, described in Nason and Silverman (1995) addresses this issue.

Under the NDWT, wavelet coefficients are calculated for all possible locations, consequently for a discrete time series of length T each level will have coefficients for locations $k \in \{0, \ldots, T - 1\}$. Some of the other benefits of the NDWT are that the transform is translation invariant. If the time points of X_t are shifted in time then the coefficients will be shifted in location but will be otherwise unchanged. Under the NDWT it is also easier to relate locations in the wavelet domain to time points as the number of locations in a scale is always equal to the number of time points. An example of the NDWT applied to a time series is shown in Figure 2.6, the series is the same as in Figure 2.5(a). Again by comparing the plots in Figure 2.6 we can see how features in the series are represented by the coefficients. Comparing the NDWT coefficients with the DWT coefficients in Figure 2.5(b) we see that the NDWT



(a) Time Series of length $T = 128$ (b) NDWT Detail Coefficients.

Figure 2.6: Non Decimated Wavelet Transform of a time series.

coefficients give much more information, for example the dip in the series just before time point $u = 0.6$ corresponds to several large coefficients in Figure 2.6(b) across a range of scales.

Having T coefficients for each scale does mean that a series will be represented by up to $T \log T$ coefficients. Clearly this is an overcomplete representation which is a consequence of the non-decimated wavelet basis not being orthonormal.

Wavelet Packet Transform: The final wavelet transform we introduce, for completeness, is the wavelet packet transform introduced by Coifman and Wickerhauser (1992). Wavelet packets are a generalisation of the ordinary DWT. Recall from Figure 2.4 in the DWT algorithm the filters h and g are applied to the smooth coefficients, $c_{j,k}$, in order to calculate the smooth and detail coefficients for the next coarsest level. The wavelet packet transform has an additional step whereby the h and g filters are

also applied to the detail coefficients to produce an additional set of smooth and detail coefficients. Clearly, like the NDWT, the wavelet packet transform is overcomplete. In order to preserve the orthogonal structure of the transform Coifman and Wickerhauser (1992) proposed the best basis algorithm to identify which coefficients are best for representing the series.

2.5 Wavelets in Time Series

In the previous section we introduced various forms of the wavelet transform. In this section we describe a nonstationary time series model which is built on a wavelet basis, namely the locally stationary wavelet model. We also summarise some recent applications of this modelling approach.

2.5.1 Locally Stationary Wavelet model

Nason et al. (2000) introduced the *locally stationary wavelet* (LSW) model to model time series which have smoothly changing spectral properties. The general form of this model is summarised below.

Definition 2.3 *Let $\{X_t\}_{t=0,\ldots,T-1}$ be a univariate time series of length $T = 2^J$. Also let $W_j(u)$ be a transfer function and $\{\xi_{jk}\}$ be a set of independent standard Gaussian increments. Finally let $\{\psi_{jk}(t)\}$ be the set of discrete nondecimated wavelets. The series X_t can then be represented as,*

$$X_t = \sum_{j=1}^{\infty} \sum_{k} W_j(k/T)\psi_{jk}(t)\xi_{jk}. \tag{2.21}$$

The transfer function, $W_j(u)$, is a Lipschitz continuous function which controls the contribution each wavelet decomposition level makes to the overall variance of the series at a particular rescaled time point u. It therefore controls the variance and autocovariance properties of the series. Other assumptions, which restrict the transfer function to be finite, are are also made, these are detailed in Definition 2.1 of Nason et al. (2000).

Nason et al. (2000) establish that the autocovariance structure of a series can be represented uniquely (up to a choice of wavelet function) in terms of the evolutionary wavelet spectrum (EWS). The EWS for level j and rescaled time point z is defined in terms of the transfer function as follows,

$$S_j(u) = |W_j(u)|^2 . \tag{2.22}$$

The property of uniqueness makes the EWS a useful quantity for analysing the autocovariance properties of the series. The autocovariance function of the series can be written explicitly in terms of the EWS in a similar way to the Fourier spectrum in equation (2.2). The nonstationary nature of the EWS means that we are no longer dealing with the global autocovariance function but rather the local autocovariance (LACV) function defined as, $c(z, \tau) = \text{cov}\,(X_z, X_{z-\tau})$.

To write the EWS in terms of the LACV we must first introduce the autocovariance wavelet, $\Psi_j(\tau) = \sum_k \psi_{j,k}(0)\psi_{j,k}(\tau)$, as defined by Nason et al. (2000). The LACV can then be written as,

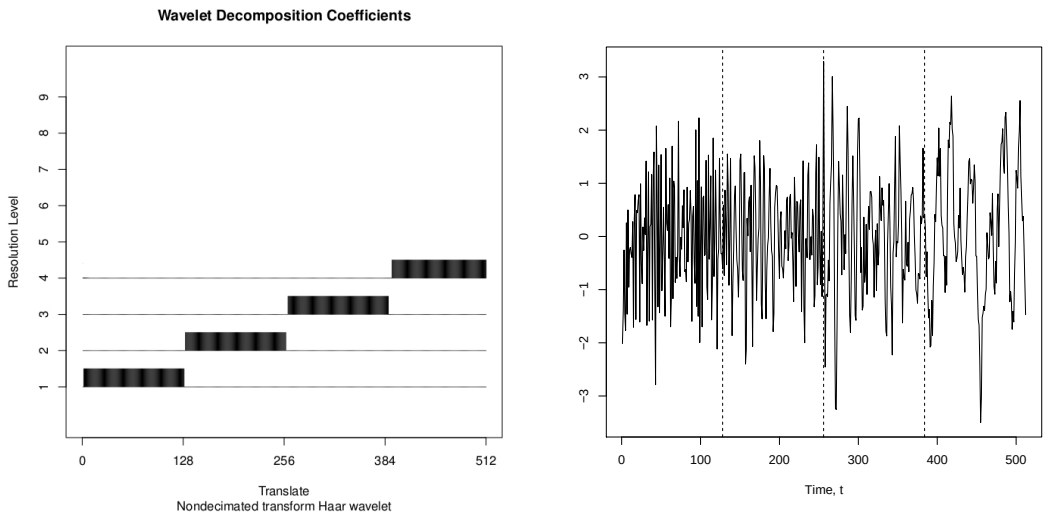
$$c(z, \tau) = \sum_{j=1}^{\infty} S_j(z)\Psi_j(\tau). \tag{2.23}$$

It is also possible to reverse this relationship and write the EWS in terms of the LACV. To do this we must also introduce the *autocorrelation wavelet inner product matrix*, $\mathbf{A}$, the j,l-th entry of this matrix is, $A_{jl} =< \Psi_j, \Psi_l >= \sum_\tau \Psi_j(\tau)\Psi_l(\tau)$. The matrix $\boldsymbol{A}$ was introduced by Nason et al. (2000) and further details about the properties of this matrix and its construction can be found in Eckley and Nason (2005). The EWS can then be expressed as,

$$S_j(z) = \sum_l A_{jl}^{-1} \sum_\tau c(z,\tau)\Psi_l(\tau). \tag{2.24}$$

As a general rule Nason et al. (2000) state that if there is high covariance between the data points X_k and $X_{k-\tau}$ then $S_j(k/T)$ should be large for a value of j which increases with τ. Intuitively this means that rapid variations in the series correspond to finer levels while slow variations correspond to coarser levels.

Example: To illustrate the role of the transfer function in controlling the autocovariance properties of the series we recreate a simulated example from Nason et al. (2000). The wavelet function used for this example is the Haar wavelet. This wavelet function is a simple step function and is shown in Figure 2.3(a). It is possible to show that a Haar LSW process where the transfer function is constant and nonzero for level j is equivalent to a moving average (MA) process of order $2^j - 1$. If the level at which the EWS is nonzero is permitted to change over time then the order of the MA process will also vary over time. For our example the EWS is chosen such that the process will initially be an MA(1) before switching to an MA(3) then MA(7) and finally MA(15). The true EWS of this process is shown in Figure 2.7(a), a simulated series with this true EWS is shown in Figure 2.7(b).



(a) An example of an EWS for a concatenation of MA processes.

(b) One realisation of this process, vertical lines indicate the transitions between processes.

Figure 2.7: An example of an EWS and one realisation of it.

Looking at Figure 2.7(b) it is easy to see how the changing EWS affects the series autocovariance. Initially the nonzero elements of the EWS are confined to the lowest level, which corresponds to high frequencies. This leads to autocovariances at short lags only which leads to rapid changes in the value of the series. Towards the end of the series the EWS becomes nonzero for higher levels, corresponding to lower frequencies. This leads to nonzero autocovariance for higher lags and consequently changes in the series happen at a much slower rate.

Estimating the EWS The procedure for estimating the EWS is similar to the procedure for estimating the Fourier spectrum. The first step is to calculate the set of wavelet detail coefficients, $d_{j,k}$, by taking a nondecimated wavelet transform of the

series as described in Section 2.4.3. The raw wavelet periodogram for level j and location k is defined as $I_k^j = |d_{j,k}|^2$.

Nason et al. (2000) establish the expectation and variance properties of the raw wavelet periodogram as,

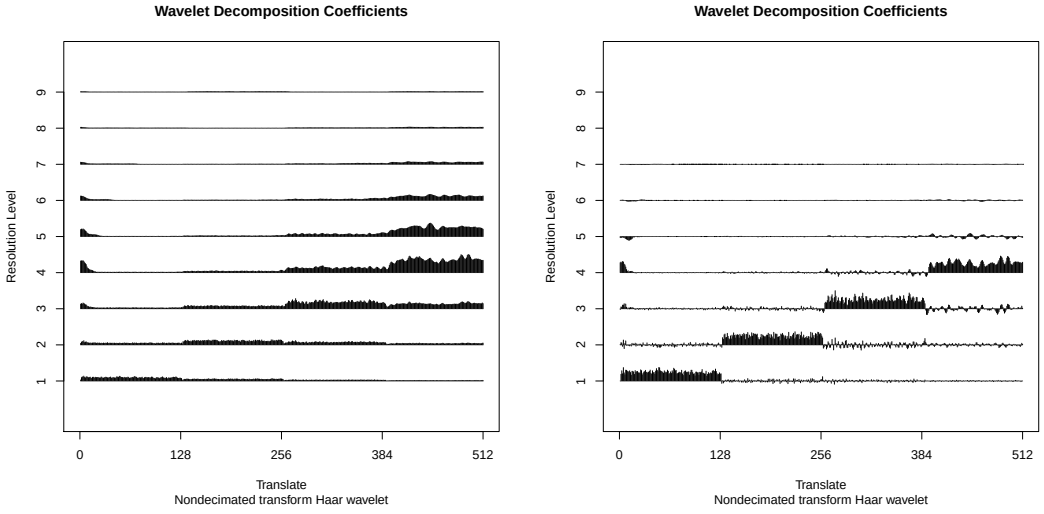
$$E\left[I^j_{[zT]}\right] = \sum_l A_{jl} S_j(z) + \mathcal{O}(T^{-1}),$$
$$\text{Var}\left\{I^j_{[zT]}\right\} = 2\left\{\sum_l A_{jl} S_l(z)\right\}^2 + \mathcal{O}(2^j/T).$$

Here A_{jl} are the elements of the autocorrelation wavelet inner product matrix defined previously. Looking at these properties we see that the raw wavelet periodogram is both biased and inconsistent. Nason et al. (2000) show that asymptotic consistency can be achieved by smoothing the raw wavelet periodogram. To this end various smoothing methods can be used. For example Nason et al. (2000) choose to use non-linear wavelet shrinkage, other methods for smoothing the wavelet periodogram can be found in Fryzlewicz and Nason (2006) and Fryzlewicz (2008).

The periodogram bias can be corrected using the inverse of $\boldsymbol{A}$. The corrected periodogram is therefore defined as, $L_k^j = \sum_l A_{jl}^{-1} I_k^l$. It is simple to show that this corrected periodogram is an unbiased estimator of the EWS. In principle the smoothing and basis correction steps can be applied in either order and the asymptotic properties of the final estimator will not be affected. Nason et al. (2000) suggest applying the smoothing step first as the distributional properties of the raw wavelet periodogram are well understood and so can be used to aid the choice of smoother.

The properties of the raw and corrected periodogram are illustrated in Figure 2.8. Here we have simulated 100 series from the true EWS shown in Figure 2.7(a). Figure

2.8(a) shows the mean of the 100 raw periodogram calculated from the 100 simulated series. It is easy to see the effect of the bias of the estimator. For example at the end of the series the only true nonzero power is in level $j = 4$ however the periodogram is clearly nonzero for levels $j = 3$ and 5. Figure 2.8(b) shows the mean of 100 corrected periodograms calculated from the same 100 simulated series. It is clear that the correction has removed the bias which was present in the raw periodogram and the resulting estimate is much closer to the true EWS.



(a) Mean of 100 Raw Periodograms (b) Mean of 100 Corrected Periodogram

Figure 2.8: An example of correcting the Raw Periodogram to reduce power leakage.

Wavelet Coherence Building upon the univariate LSW framework Sanderson et al. (2010) introduced a bivariate extension which includes a first definition of LSW co-

herence. The two channels of a bivariate LSW series are represented as follows,

$$X_t^1 = \sum_{j=1}^{\infty} \sum_{k=-\infty}^{\infty} W_j^{(1)}(k/T)\psi_{j,t-k}\xi_{j,k}^{(1)},$$
$$X_t^2 = \sum_{j=1}^{\infty} \sum_{k=-\infty}^{\infty} W_j^{(2)}(k/T)\psi_{j,t-k}\xi_{j,k}^{(2)}$$

It is easy to see that when viewed individually each channel of the series has the same form as the univariate LSW model defined in equation (2.21). The main difference with the bivariate LSW comes via the set of random innovations $\{\xi_{j,k}^{(1)}\}$ and $\{\xi_{j,k}^{(2)}\}$. Individually both sets of innovations have the same distributional properties as those in equation (2.21). However, in addition Sanderson et al. (2010) also require that they have the following covariance properties: $\text{cov}\left(\xi_{j,k}^{(i)}, \xi_{j',k'}^{(i)}\right) = \delta_{j,j'}\delta_{k,k'}$ and $\text{cov}\left(\xi_{j,k}^{(1)}, \xi_{j',k'}^{(2)}\right) = \delta_{j,j'}\delta_{k,k'}\rho_j(k/T)$, thereby introducing a dependence structure between the bivariate signal components. Sanderson et al. (2010) call this dependence the coherence.

The autocovariance of each channel of the series is characterised by its own EWS, $S_j^{(1)}(u)$ and $S_j^{(2)}(u)$. Additionally the covariance between the channels is characterised by the evolutionary wavelet cross-spectrum, $C_j(z)$, which is defined as,

$$C_j(u) = W_j^{(1)}(u)W_j^{(2)}(u)\rho_j(u).$$

The cross covariance between the two channels at rescaled time point u and lag τ is denoted as $c^{(1,2)}(u,\tau)$. This is related to the cross-spectrum as,

$$c^{(1,2)}(u,\tau) = \sum_{j=1}^{\infty} C_j(u)\Psi_j(\tau),$$

mirroring the relationship between the EWS and the autocovariance in the univariate setting.

The cross spectrum can be estimated in a similar way to the EWS. Sanderson et al. (2010) define the raw cross-periodogram as, $I_{j,t}^{(1,2)} = d_{j,t}^{(1)} d_{j,t}^{(2)}$. They show that the cross-periodogram has the following properties,

$$\begin{aligned} E\left[I_{j,t}^{(1,2)}\right] &= \sum_{l=1}^{\infty} A_{jl} W_l^{(1)}(t/T) W_l^{(2)}(t/T) \rho_l(t/T) + 2^j O(T^{-1}), \\ \mathrm{Var}\left\{I_{j,t}^{(1,2)}\right\} &= \sum_{l=1}^{\infty} A_{jl} S_l^{(1)}(t/T) \sum_{l'=1}^{\infty} A_{jl'} S_{l'}^{(2)}(t/T), \\ &\quad + \left\{\sum_{i=1}^{\infty} A_{jl} W_l^{(1)}(t/T) W_l^{(2)}(t/T) \rho_i(t/T)\right\}^2 + 2^j O(T^{-1}). \end{aligned}$$

The raw cross-periodogram therefore suffers from the same problems of bias and inconsistency as the (univariate LSW) raw periodogram. This can be overcome by smoothing and correcting to produce an asymptotically unbiased and consistent estimate.

The wavelet coherence can be written in terms of the two EWS' and the cross spectrum,

$$\rho_j(u) = \frac{C_j(u)}{\left[S_j^{(1)}(u) S_j^{(2)}(u)\right]^{\frac{1}{2}}}. \tag{2.25}$$

The wavelet coherence takes a value on the interval, $[-1, 1]$. This is distinct from the Fourier coherence which is defined to be always positive. As such the wavelet coherence is more similar to Fourier coherency however for the remainder of this thesis we will adopt the convention of Sanderson et al. (2010) and refer to $\rho_j(u)$ as wavelet coherence.

The wavelet coherence can be estimated using the estimated EWS for the two

channels and the estimated cross-spectrum as follows,

$$\widehat{\rho}_j(t/T) = \frac{\widehat{C}_j(t/T)}{\left[\widehat{S}_j^{(1)}(t/T)\widehat{S}_j^{(2)}(t/T)\right]^{\frac{1}{2}}}. \tag{2.26}$$

The method of Sanderson et al. (2010) is restricted to only cover bivariate time series and as such cannot be used to answer questions relating to more than two components such as those raised at the end of Section 2.2.2. Their particular construction of the model also requires information about the second order structure to be encoded separately in the random elements and the transfer functions.

2.5.2 Applications of the LSW Model

We conclude our review of the LSW approach by summarising recent developments in the LSW literature during the last few years. In particular we focus on the key developments which have been made in the areas of LSW forecasting, test of stationarity, classification and changepoint detection.

Forecasting: The first application of the LSW model which we discuss is the Forecasting of time series. The use of the LSW model for forecasting was introduced by Fryzlewicz et al. (2003). They point to the problem of predicting future observations for nonstationary series where there is only a short sections of homogeneous structure at the end of the observed series. The LSW model is suited for such a scenario as it does not assume stationary in the autocovariance structure.

The forecasting procedure which they propose is based on a linear predictor. Given the LSW series X_t which is observed at time points $t \in \{0, 1, \ldots, T-1\}$, Fryzlewicz

et al. (2003) define the h-step ahead predictor as,

$$\hat{X}_{t-1+h} = \sum_{s=0}^{t-1} b_{t-1-s}^{(h)} X_t, \tag{2.27}$$

where the coefficients $b_{t-1-s;T}^{(h)}$ are chosen to minimise the mean square prediction error: $\text{MSPE}\left(\hat{X}_{t-1+h}, X_{t-1+h}\right) = \text{E}\left[\hat{X}_{t-1+h} - X_{t-1+h}\right]^2$. They establish that the MSPE of the one step ahead forecast can be expressed in terms of the vector of coefficients $\boldsymbol{b}_t = [b_{t-1}^{(1)}, \ldots, b_0^{(1)}, -1]$ and the matrix $\boldsymbol{\Sigma}_t$ which is the covariance matrix of $X_{0,T}, \ldots, X_{t-1}$. The MSPE is then given by: $\text{MSPE}\left(\hat{X}_t, X_t\right) = \boldsymbol{b}_t' \boldsymbol{\Sigma}_t \boldsymbol{b}_t$. In order to apply this prediction to real data the coefficients $\boldsymbol{b}_t$ must be estimated.

Since the true covariance matrix is not known, Fryzlewicz et al. (2003) estimate it using the matrix $\boldsymbol{B}_t$, the (m,n)-th element of which is given by: $\sum_{j=1}^{J} S_j\left(\frac{m+n}{2T}\right) \Psi_j(n-m)$, where $S_j(u)$ is the EWS for X_t which can be estimated using the methods described in Section 2.5.1. Fryzlewicz et al. (2003) then show, under some assumptions on the covariances and the spectrum, that $\text{MSPE}\left(\hat{X}_t, X_t\right) = \boldsymbol{b}_t' \boldsymbol{B}_t \boldsymbol{b}_t (1 + \mathcal{O}_T(1))$. They also establish that the set of coefficients, $\{b_s^{(1)}\}$, which minimises the MSPE must be the solution to the following set of linear equation,

$$\sum_{m=0}^{t-1} b_{t-1-m;T}^{(1)} \left\{ \sum_{j=1}^{J} s\left(\frac{m+n}{2T}\right) \Psi_j(m-n) \right\} = \sum_{1}^{J} S_j\left(\frac{t+n}{2T}\right) \Psi_j(t-n).$$

It is possible to invert this system of equations and thus calculate the coefficients needed to make a one step ahead prediction. A more general h-step ahead prediction is also covered by Fryzlewicz et al. (2003).

A Test of Stationarity: As has been discussed previously it is important that any nonstationarity is taken into account when analysing a time series. It is therefore

desirable to be able to test a series to establish whether or not it is stationary. Such a test was introduced by Nason (2013). The test makes use of the LSW model and is able to not only identify if the series is nonstationary but also identify the locations at which such points of nonstationary behaviour occur. Their method has been implemented in `R` using the `locits` package. For a time series X_t with true EWS $S_j(u)$ Nason (2013) define the function $\beta_j(u)$ as,

$$\beta_j(u) = \sum_{l=1}^{\infty} A_{lj} S_l(u). \tag{2.28}$$

As shown in Section 2.5.1 this is the expected value of the raw wavelet periodogram, $I_{j,uT}$. Clearly if X_t is a second order stationary series then $S_j(u)$ and therefore $\beta_j(u)$ will be constant over time for all values of j. It was noted by von Sachs and Neumann (2000) that if $\beta_l(u)$ is constant for all values of u then its Haar wavelet coefficients, $v_{i,p}^{(l)}$ will be zero for all values of i and p. These coefficients are calculated as: $v_{i,p}^{(l)} = \int_0^1 \beta_l(u)\psi_{i,p}^H(u)\,\mathrm{d}u$, where $\psi_{i,p}^H(u)$ is the Haar wavelet function shown in Figure 2.3(a). The true function $\beta_j(u)$ is generally not known and so is replaced by its estimate, $I_{j,uT}$, to give the estimated Haar coefficients $\hat{v}_{i,p}^{(j)}$.

The test statistics for the test of stationarity is chosen to be, $T_{i,p}^{(l)} = \hat{v}_{i,p}^{(l)}\hat{\sigma}_{i,p}^{(l)-1}$, where $\hat{\sigma}_{i,p}^{(l)}$ is the estimated standard deviation of the Haar wavelet coefficients. Nason (2013) show that for a series of length T this standard deviation is estimated as,

$$\hat{\sigma}_{i,p}^{(l)2} = 2T^{-1} I_{l,\langle 1,T\rangle}^2 \int_0^1 \psi_{i,p}^H(u)^2\,\mathrm{d}u = 2T^{-1} I_{l,\langle 1,T\rangle}^2,$$

where, $I_{l,\langle 1,T\rangle}^2 = T^{-1}\sum_t I_{l,t}^2$. Under the null hypothesis, H_0, the coefficient, $v_{i,p}^{(l)}$, is zero and the test statistic, $T_{i,p}^{(l)}$, follows a standard normal distribution. The null hypothesis can therefore be evaluated by comparing the test statistic to a critical value in the

usual way. Since there are many different Haar coefficients which need to be tested this is a multiple hypothesis test, Nason (2013) suggest using either Bonferroni correction or the false discovery rate, FDR, procedure of Benjamini and Hochberg (1995) to control the rate of false positives. If the null hypothesis is rejected for coefficient $v_{i,p}^{(l)}$ then the values of i, p and l will yield information about the region in the time and frequency decomposition of X_t which contains the nonstationarity.

Time Series Classification: The next application which we will describe is the classification of time series. We will first describe the LSW based method of Fryzlewicz and Ombao (2009) and then describe extensions to this method found in Krzemieniewska et al. (2014). We assume that a time series X_t will belong to one of G different classes where G is known. It is assumed that all series which belong to the same class will have the same underlying LSW process, in other words a series which belongs to class Π_g will have a LSW representations with EWS $S_j^{(g)}(u)$.

To estimate the EWS for a particular class Fryzlewicz and Ombao (2009) assume that a set of training data exists such that there are N_g independent series belonging to class Π_g. The corrected wavelet periodogram is then calculated for each of the series in the training set, with the n-th series belonging to class Π_g denoted as $L_{j,k}^{g,n}$. These periodograms are then used to estimate the spectrum as follows, $\hat{S}_j^{(g)}(k/T) = {N_g}^{-1} \sum_{n=1}^{N_g} L_{j,k}^{g,n}$.

To better distinguish between classes Fryzlewicz and Ombao (2009) suggest using a subset of the coefficients of the EWS denoted as $\mathcal{M}$. This subset is chosen based on a divergence measure. For the case of $G = 2$ this is defined as,

$\Delta(j,k) = \left[S_j^{(1)}(k/T) - S_j^{(2)}(k/T)\right]^2$. The divergence measures is calculated for all possible values of j and k and ordered. The subset $\mathcal{M}$ contains a prespecified proportion of timescale indices which have the highest divergence values.

A series with unknown class membership, X_t, can be classified by computing the squared quadratic distance, D_g, between its corrected periodogram, $L_{j,k}$, and the estimated EWS for each class, $D_g = \sum_{(j,k)\in\mathcal{M}}\{L_{j,k} - \hat{S}_{j,k}\}^2$. The series is therefore classified in the class corresponding to the lowest value of D_g.

As an extension to this work Krzemieniewska et al. (2014) note that the coefficients with the highest divergence may not necessarily produce the most consistent classification. They suggest an alternative divergence measure $\tilde{\Delta}(j,k)$ which is related to $\Delta(j,k)$ by the formula,

$$\tilde{\Delta}(j,k) = \Delta(j,k)/\hat{\sigma}^2_{j,k},$$

where $\hat{\sigma}^2_{j,k}$ is the variance of $L_{j,k}$. The idea being the most divergent and stable coefficients are used in a classification approach. In addition to this new divergence measure Krzemieniewska et al. (2014) also modify the distance measure such that: $\tilde{D}_g = \sum_{(j,k)\in\mathcal{M}} \frac{\left\{L_{j,k} - \hat{S}_j^{(g)}(k/T)\right\}^2}{\hat{\sigma}^2_{j,k}}$. In simulation studies they demonstrate that these modifications lead to an increase in classification accuracy compared to Fryzlewicz and Ombao (2009)

Changepoint Detection: The final application we focus on is the detection of changpoints in the autocovariance structure of a time series. This problems has previously been studied by Davis et al. (2006) and Gombay (2008) but we will focus on the LSW based method of Killick et al. (2013). Their method assumes that a

series $\mathbf{X}_t = \{X_0, \ldots, X_{n-1}\}$ is Gaussian. This assumption is valid if the set of random innovations $\{\xi_{j,k}\}$ from equation (2.21) are Gaussian. The authors begin by defining a hypothesis test for a single autocovariance changepoint. The null hypothesis, H_0, and alternative, H_1, are defined as follows,

$$\begin{aligned} H_0 : &\operatorname{cov}(X_0, X_{0-\nu}) = \operatorname{cov}(X_1, X_{1-\nu}) = \ldots = \operatorname{cov}(X_{n-1}, X_{n-1-\nu}) = c_{0,\nu}, \ \forall \nu \geq 0, \\ H_1 : &c_{1,\nu} = \operatorname{cov}(X_0, X_{0-\nu}) = \ldots = \operatorname{cov}(X_\tau, X_{\tau-\nu}) \\ &\neq \operatorname{cov}(X_{\tau+1}, X_{\tau+1-\nu}) = \ldots = \operatorname{cov}(X_{n-1}, X_{n-1-\nu}) = c_{n,\nu}, \ \forall \nu \geq 0, \end{aligned} \quad (2.29)$$

where ν is the autocovariance lag. The null hypothesis is equivalent to the series being second order stationary. In the LSW representation of X_t a second order stationary series will have an EWS which is constant over time and so, $W_j^2(k/n) = \gamma_j$ at every scale j. The alternative hypothesis is equivalent the the series being split into two second order stationary segments. Clearly the hypotheses defined in equation (2.29) can be written as,

$$\begin{aligned} H_0 :& W_j^2\left(\frac{0}{n}\right) = W_j^2\left(\frac{1}{n}\right) = \ldots = W_j^2\left(\frac{n-1}{n}\right) = \gamma_{0,j}, \quad \forall j, \\ H_1 :& \gamma_{1,j} = W_j^2\left(\frac{0}{n}\right) = \ldots = W_j^2\left(\frac{\tau}{n}\right) \neq W_j^2\left(\frac{\tau+1}{n}\right) = \ldots = W_j^2\left(\frac{n-1}{n}\right) = \gamma_{n,j}, \end{aligned} \quad (2.30)$$

for some $j \in \{1, 2, \ldots\}$.

In order to perform the hypothesis test Killick et al. (2013) express the likelihood of a Gaussian LSW series in terms of the transfer function. Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be observations of an LSW process with Gaussian innovations. The log-likelihood for this series can be expressed as: $\ell(W|\mathbf{x}) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_W| - \frac{1}{2}\mathbf{x}'\Sigma_W^{-1}\mathbf{x}$. Consequently

using equation (2.23) the elements of the variance covariance matrix, Σ_W, can be expressed as,

$$\Sigma_W(k,k') = \text{cov}\,(X_k, X_{k'}) = \sum_{l,m} W_l^2 \left(\frac{m}{n}\right) \psi_{l,m-k}\psi_{l,m-k'}.$$

Using this form of the log-likelihood Killick et al. (2013) define a likelihood ratio test statistic for the single changepoint test as,

$$\lambda = \max_{J<\tau<n-J} \left\{ \log\left|\hat{\Sigma}_0\right| + \mathbf{x}'\hat{\Sigma}_0^{-1}\mathbf{x} - \log\left|\hat{\Sigma}_1\right| - \mathbf{x}'\hat{\Sigma}_1^{-1}\mathbf{x} \right\}.$$

Here $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$ are the maximum likelihood estimates of the variance covariance matrix under the null and alternative hypotheses respectively. These estimates are defined as,

$$\begin{aligned}
\hat{\Sigma}_0(k,k') &= \sum_l \sum_m \hat{\gamma}_{0,l}\psi_{l,m-k}\psi_{l,m-k'}, \\
\hat{\Sigma}_1(k,k') &= \sum_l \left[\sum_{m\leq\tau} \hat{\gamma}_{1,l}\psi_{l,m-k}\psi_{l,m-k'} + \sum_{m>\tau} \hat{\gamma}_{n,l}\psi_{l,m-k}\psi_{l,m-k'} \right].
\end{aligned}$$

Using the above test statistic a changepoint is deemed significant if $\lambda > c$ for some pre-defined constant c.

The test described above can be extended to a multiple changepoint setting using various search algorithms such as Binary Segmentation introduced by Scott and Knott (1974).

Chapter 3

Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes

Abstract

We consider the problem of estimating time-localized cross-dependence in a collection of non-stationary signals. To this end we develop the multivariate locally stationary wavelet framework which provides a time-scale decomposition of the signals and thus naturally captures the time-evolving scale-specific cross-dependence between components of the signals. Under the proposed model, we rigorously define and estimate two forms of cross-dependence measures: wavelet coherence and wavelet partial coherence. These dependence measures differ in a subtle but important way. The former is a broad measure of dependence which may include indirect associations, i.e. dependence between a pair of signals that is driven by another signal. Conversely, wavelet partial coherence measures direct linear association between a pair of signals, i.e. it removes the linear effect of other observed signals. Our time-scale wavelet partial coherence estimation scheme thus provides a mechanism for identifying hidden dynamic relationships within a network of non-stationary signals, as we demonstrate on electroencephalograms recorded in a visual-motor experiment.

3.1 Introduction

Historically much of the literature on non-stationary signals is focused on the univariate setting. For reviews of this area see Cohen (1989); Dahlhaus (2012); Daubechies (1990); Kayhan et al. (1994); Kumar and Fuhrmann (1992); Priestley (1988) and references therein. However with advanced data collection devices such as those used in the medical and mobile sectors, there is a need for rigorous approaches to assess and confirm time-localized direct vs. indirect dependence (or lack thereof) between signals. It is often difficult to infer dynamic cross-dependence between components of multivariate signals such as the multi-channel EEG collected during a visual-motor task (see Figure 3.1) which we will revisit later.

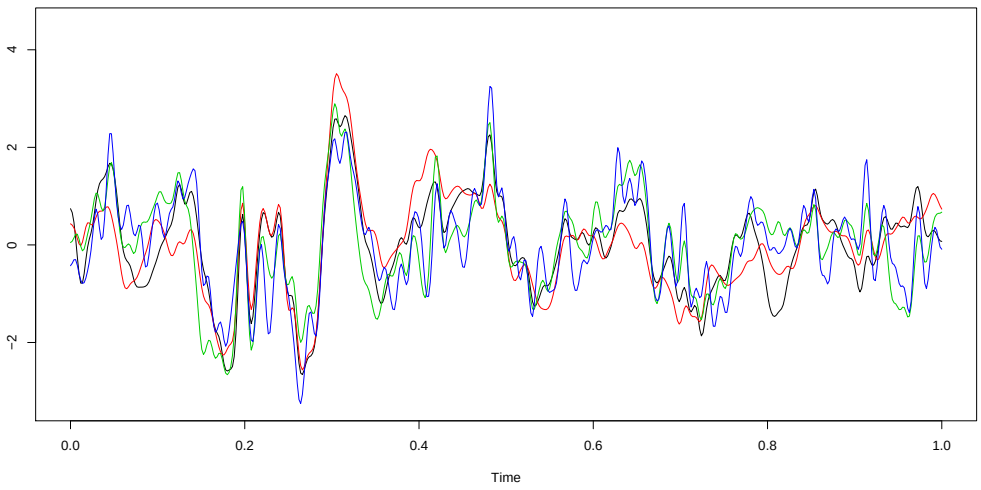


Figure 3.1: Plot of a 4-channel EEG.

We consider precisely this challenge, developing a novel approach for characterizing and estimating cross-dependence between non-stationary signals having dynamic and complex cross-dependence structures. In doing so, we highlight two specific forms of

dependence which can be estimated between pairs of signals within a multivariate collection. The simplest form is that of the (time-dependent) coherence between two signals. This describes the linear relationship between two signals - more precisely it is a time-evolving squared cross-correlation between filtered signals, Ombao and Van Bellegem (2008). However, in so doing we may also include indirect associations driven by another observed signal in the collection. The alternative is partial coherence. This provides a measure of the direct linear relationship between two signals over time, thus removing the (linear) effects of other observed signals. The difference between direct vs indirect associations is illustrated in Figure 3.2. This measure has broad potential scientific impact, for example the the neuroscience and genomic communities are keenly interested in such associations.

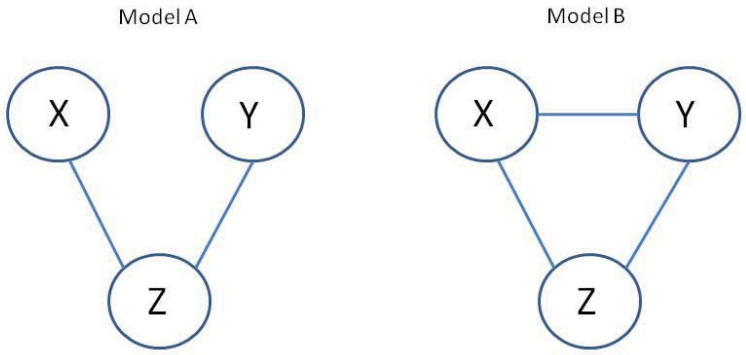


Figure 3.2: Indirect vs. Direct Associations Between Signals. Left: X and Y are indirectly linked through Z. Right: X and Y are directly linked. Coherence between X and Y is non-zero for both networks. Partial coherence is non-zero for the network on the right (with direct link) but zero for the left network because the link between X and Y is indirect.

Previous Work In recent years, several papers have appeared trying to address the non-stationary modelling challenge associated with such large and complex signals. In Dahlhaus (2000a), Dahlhaus presents a Fourier based model for multivariate locally stationary signals with time-varying spectral structure. A similar approach was also developed by Walden and Cohen (2012). Under the Dahlhaus framework, Ombao and Van Bellegem Ombao and Van Bellegem (2008) demonstrate that the time-varying coherence is equivalent to the modulus-squared cross-correlation between filtered segmented signals. Segment sizes are obtained data-adaptively by iteratively increasing segment lengths as long as the stationarity assumption within each segment is not violated. Such a data-adaptive windowing approach, however, is computationally demanding. An alternative Fourier based approach to model multivariate non-stationary series is the smooth localized complex exponential (SLEX) model of Ombao et al. Ombao et al. (2005). Here the best representation of the signal is selected from the SLEX library using a complexity-penalized Kullback-Leibler criterion. Although capable of handling massive signals, the SLEX method is restricted to choosing representations obtained from temporally-dyadic segmentation. Moreover we note that both Ombao and Van Bellegem (2008) and Ombao et al. (2005) only develop methods for the estimation of coherence which, as we shall show later, can mask understanding of the direct relationships between pairs of signal components.

Cohen and Walden Cohen and Walden (2010) overcome the limitations of dyadic temporal splits within SLEX by using a wavelet basis to adapt to nonstationarity in the spectra of each channel for the case of jointly stationary processes. The assumption of jointly stationary processes is not present in Cohen and Walden (2011) and

Sanderson et al. (2010) who both use wavelet based models to quantify non-stationary linear dependence between components of a *bivariate* non-stationary signals. More recently, within the more restricted context of changepoint detection of piecewise stationary signals, Cho and Fryzlewicz (2014) has extended the approach of Sanderson et al. (2010) to a p-variate setting. However none of these contributions directly address the issues that are germane to truly multivariate non-stationary signals (with three or more components). Specifically, as Koopmans (1964) identified in the stationary context, one major practical issue is to identify whether the (time-dependent) connection or cross-dependence between two channels is either (a.) direct or (b.) indirect (i.e., driven by another channel or common set of channels). It is this challenge which lies at the heart of this article.

Our Work: The modelling framework which we propose in this paper is an alternative formulation of the model form proposed by Sanderson et al. (2010). The model proposed by Sanderson et al. (2010) decomposes the spectral and cross-spectral structure into two different components: the within-channel structure being encapsulated within the transfer functions whilst the cross-channel structure is contained within the process innovations. Instead we propose a more parsimonious form, whereby both spectral components are described within a matrix of transfer functions. Specifically, to extract cross-dependence structures, we introduce the multivariate locally stationary wavelet framework (MvLSW) - which is a stochastic representation that is ideally suited for non-stationary signals. This framework permits the direct estimation of both the coherence *and* partial coherence in a computationally efficient manner. In addition the framework also permits direct simulation of processes with a specific

time-scale partial coherence form, including processes with abrupt changes in partial coherence. This direct simulation is necessary to perform resampling-based inference.

The format of the rest of the paper is as follows. Our main contributions are developed in Sections 3.2 and 3.3. Specifically, in Section 3.2.1 we develop the multivariate locally stationary wavelet framework for modelling multivariate signals. We then introduce the local wavelet spectral matrix as a representation of the properties of the signals in Section 3.2.2. In Section 3.2.3 we use the MvLSW model to develop our two key cross-dependence quantities: wavelet coherence and partial coherence. Section 3.3 gives detail of the estimator for the local wavelet spectral matrix as well as establishing its asymptotic properties. Finally Section 3.4 provides an example of how our approach can be used to identify direct time-dependent relationships between components of a signal which we demonstrate on multi-channel electroencephalograms (EEGs) recorded during a visual-motor experiment, as well as on simulated data.

3.2 Locally Stationary Wavelet Processes

This section describes the multivariate LSW (MvLSW) modelling framework, together with various time-scale measures which we introduce to describe the spectral and cross-spectral behaviour of such non-stationary signals. For completeness we start by briefly reminding the reader of key aspects associated with *univariate* LSW theory as introduced by Nason et al. (2000), their building blocks (discrete wavelets) and the associated evolutionary wavelet spectrum (EWS).

The key building blocks in constructing LSW processes, discrete wavelets, are

founded on $\{h_k\}$ and $\{g_k\}$, the usual low and high-pass quadrature mirror filters associated with the construction of Daubechies' compactly supported continuous-time wavelets. The associated **discrete wavelets**, $\psi_j = \{\psi_{j,0}, \psi_{j,1}, \ldots, \psi_{j,N_j-1}\}$ are vectors of length N_j for scales $j \in \mathbb{N}$ which can be calculated using the following: $\psi_{1,n} = \sum_k g_{n-2k}\delta_{0,k} = g_n$ for $n = 0, \ldots, N_1 - 1$ and $\psi_{j+1,n} = \sum_k h_{n-2k}\psi_{j,k}$, for $n = 0, \ldots, N_{j+1} - 1$. Here $\delta_{0,k}$ is the usual Kronecker-delta function, and $N_j = (2^j - 1)(N_h - 1) + 1$ where N_h is the number of non-zero elements within the filter $\{h_k\}$. The discrete wavelets form the corner-stone of the (univariate) LSW time series model. Specifically, assume that $T = 2^J$ for some $J \in \mathbb{Z}$. Then the LSW process, $X_{t;T}$, is defined to be a sequence of (doubly-indexed) stochastic processes having the following representation in the mean-square sense:

$$X_{t;T} = \sum_{j=1}^{\infty} \sum_{k} W_j(k/T)\psi_{j,t-k}\xi_{j,k}. \tag{3.1}$$

As described in Nason et al. (2000), the representation consists of the discrete wavelets; $\{W_j(u)\}_{u\in(0,1)}$, a smoothly varying transfer function and $\{\xi_{j,k}\}$, a collection of zero-mean, unit-variance uncorrelated random variables. A number of smoothness assumptions are also required on the $\{W_j(\cdot)\}$ to ensure that the transfer function can be estimated (see Nason et al. (2000) for details).

The transfer function, $W_j(k/T)$, provides a measure of the time-varying contribution to the variance at a particular scale, j. Consequently, to describe the power contained at a given scale and location, Nason et al. (2000) introduce the evolutionary wavelet spectrum (EWS), $S_j(u) = |W_j(u)|^2$, for $j \in \mathbb{N}$. This can be estimated using the wavelet periodogram for a one-dimensional non-stationary signal, see Nason et al.

(2000) for details.

3.2.1 The Multivariate LSW model

We now introduce our multivariate generalization of the LSW framework. In what follows we will refer to each (univariate) component signal as a channel. Our main goal is to develop a framework for modeling multivariate non-stationary signals under which we rigorously define the time-varying second order properties, and in particular the locally stationary cross-dependence between the different channels. In our framework we allow individual channels to experience their own uniquely localized non-stationary behaviour. More importantly we explicitly describe the potentially locally stationary correlation between channels. Under our model this correlation will be broken down into contributions from different scales. This is known as the coherence structure. It is important to be able to represent this structure adequately as it will reveal how the channels relate to each other and how this can change over time.

We start by considering a P-dimensional vector, $\mathbf{X}_{t;T} = [X^{(1)}_{t;T}, X^{(2)}_{t;T}, \ldots, X^{(P)}_{t;T}]'$, each element of which is an individual channel of the signal. To represent this signal under a multivariate model we replace the transfer function, $W_j(k/T)$, from the (univariate) LSW model with a $P \times P$ matrix of functions, $\mathbf{V}_j(k/T)$, known as the *transfer function matrix*. The innovations, $\{\xi_{jk}\}$, are also replaced by a set of random vectors, $\{\mathbf{z}_{j,k}\} = \{[z^{(1)}_{j,k}, \ldots, z^{(P)}_{j,k}]'\}$. The definition of the *multivariate LSW model* is then given as follows.

Definition 3.1 *The P-variate locally stationary wavelet process* $\{\mathbf{X}_{t;T}\}_{\{t=0,\ldots,T-1\}}$,

$T = 2^J$, $J \in \mathbb{N}$ *is represented by,*

$$\mathbf{X}_{t;T} = \sum_{j=1}^{\infty} \sum_{k} \mathbf{V}_j(k/T)\psi_{j,t-k}\mathbf{z}_{j,k}, \tag{3.2}$$

where $\{\psi_{j,t-k}\}_{jk}$ *is a set of discrete non-decimated wavelets;* $\mathbf{V}_j(k/T)$ *is the transfer function matrix, which is defined to have a lower-triangular form. We assume that each element of the transfer function matrix is a Lipschitz continuous function with Lipschitz constants* L_j *satisfying* $\sum_{j=1}^{\infty} 2^j L_j^{(p,q)} < \infty$*;* $\mathbf{z}_{j,k}$ *are uncorrelated random vectors with mean vector* $\mathbf{0}$ *and variance-covariance matrix equal to the* $P \times P$ *identity matrix.*

We will henceforth drop the explicit dependence of the process on T, although naturally it will still be assumed.

Remark. The distributional property of the random elements in Definition 3.1 means that the elements have the following covariance property: $\text{cov}\left(z_{j,k}^{(i)}, z_{j',k'}^{(i')}\right) = \delta_{i,i'}\delta_{j,j'}\delta_{k,k'}$. In other words the $\{z_{j,k}^{(i)}\}$ are random orthonormal increment sequences, which are themselves uncorrelated. Dependence between channels is encapsulated *only* in the transfer function matrix which also controls the contribution to the variance made by each channel at a particular time within each scale. This differs from the approach in Sanderson et al. (2010) where the dependence structure is encapsulated within the innovations $\mathbf{z}$.

Remark. The primary difference between our approach and that of Sanderson et al. (2010), or indeed the more recent contribution of Cho and Fryzlewicz (2014), is that in our framework we encapsulate the spectral structure (including cross-channel dependence) entirely within the transfer function matrix. This is in contrast to the

Sanderson *et al.* framework, where the spectral structure is encapsulated both within (i) the transfer functions (spectrum) and (ii) process innovations (cross-channel dependence). As such our framework permits one to estimate the *partial* coherence in a straightforward manner, since this structure is entirely embedded within the transfer function matrix. Computationally there are also benefits to this particular formulation: for example, this approach can be implemented via matrix operations, whilst in the formulation of [14] one would conduct the estimation scheme on each channel individually. More importantly, perhaps, it is possible to simulate multivariate time series with a given partial coherence form directly within this framework. The ability to perform such simulations means that resampling based inference can be performed in this setting.

Many different forms of transfer function matrix could be chosen, however for ease of interpretation we choose for it to have a lower triangular form. The lower triangular form of $\mathbf{V}_j(u)$ makes it very easy to generalize to multiple dimensions. It is also easy to see how linear dependencies between the channels are produced. If the off diagonal terms are non-zero then there will be (time-varying) dependence between the series, however if $\mathbf{V}_j(u)$ is diagonal then the channels will be uncorrelated with each other. Here, we do not estimate $\mathbf{V}_j(u)$ but estimate the spectral quantities which we discuss in the next subsection. Moreover the lower triangular form can represent a general spectral structure even if the channel order is permuted. This is explained further in Proposition 3.3.

3.2.2 Local Wavelet Spectral and Covariance Matrices of Non-Stationary signals

We next introduce the local wavelet spectral matrix which describes the time-scale decomposition of power in our multivariate time series. Recall that in the univariate LSW context the concept of an evolutionary wavelet spectrum describes a time-scale decomposition of power. Since we are dealing with multivariate signals, and have replaced the transfer function with a transfer function matrix, we will introduce its multivariate analog – the *local wavelet spectral matrix*.

Definition 3.2 *Let $\mathbf{X}_t$ be a MvLSW signal with associated time-dependent transfer function matrix $\mathbf{V}_j(u)$. Then the local wavelet spectral (LWS) matrix at scale j and rescaled time u is defined to be,*

$$\mathbf{S}_j(u) = \mathbf{V}_j(u)\mathbf{V}_j'(u), \tag{3.3}$$

where $\mathbf{V}_j'(u)$ denotes the transpose of $\mathbf{V}_j(u)$.

Remark. The LWS matrix provides a measure of the local contribution to both the variance of the channels and cross-covariance between channels made at a particular time, u, and scale, j. By the construction of Definition 3.2 it is clear that for any given transfer function matrix the LWS matrix is symmetric and positive semi-definite for every fixed time-scale combination. The diagonal elements of the LWS matrix are the spectra of the individual channels of the signals and are denoted $S_j^{(p,p)}(u)$. The off diagonal terms, $S_j^{(p,q)}(u)$, describe the cross-spectra between the series. It is also natural to consider whether a connection can be established between the LWS matrix

and the local auto and cross-covariance. We start to explore this connection in the following definition. However prior to doing so we introduce the discrete autocorrelation wavelet, $\Psi_j(\tau)$, which is defined by $\Psi_j(\tau) \equiv \sum_k \psi_{j,k}\psi_{j,k-\tau}$ for $j \in \mathbb{N}$ and $\tau \in \mathbb{Z}$ (see Eckley and Nason (2005) for further details).

Definition 3.3 *Let $c^{(p,p)}(u,\tau)$ denote the local autocovariance of channel p at lag τ and $c^{(p,q)}(u,\tau)$ be the local cross-covariance between channels p and q. We can define these function in terms of the elements of the LWS matrix and the discrete autocorrelation wavelets,*

$$c^{(p,p)}(u,\tau) = \sum_{j=1}^{\infty} S_j^{(p,p)}(u)\Psi_j(\tau),$$
$$c^{(p,q)}(u,\tau) = \sum_{j=1}^{\infty} S_j^{(p,q)}(u)\Psi_j(\tau). \tag{3.4}$$

The following proposition establishes that, up to choice of wavelet, the LWS matrix is unique for a specified MvLSW model form.

Proposition 3.1 *Given the corresponding MvLSW process, the LWS matrix is uniquely defined.*

Proof: See Appendix A.1.

We also consider if under this definition the local auto- and cross-covariance functions exactly represent the covariance between elements of the signals.

Proposition 3.2 *Let $c^{(p,q)}(u,\tau)$ denote the local cross covariance stated in Definition 3.3. This function can also be represented, approximately, in terms of the covariance between elements of the signal because*

$$\left| c^{(p,q)}(u,\tau) \; - \; \operatorname{cov}\left(X^{(p)}_{[uT]}, X^{(q)}_{[uT]+\tau} \right) \right| = \mathcal{O}(T^{-1}).$$

Proof: See Appendix A.2.

Remark. Given the lower triangular form of the transfer function matrix, $\mathbf{V}_j(u)$, it is natural to ask if the representation is reliant on a certain ordering of the channels of $\mathbf{X}_t$. It is possible to show that under any permutation of this ordering $\mathbf{X}_t$ will have a MvLSW representation and the spectral properties will be unchanged.

Proposition 3.3 *Let $\mathbf{X}_t$ be a MvLSW process with LWS matrix, $\mathbf{S}_j(u)$. Also let $\mathbf{X}_t^*$ be a permutation of $\mathbf{X}_t$ such that $\mathbf{X}_t^* = \mathbf{M}\mathbf{X}_t$ for some permutation matrix $\mathbf{M}$. Then the LWS matrix of $\mathbf{X}_t^*$, $\mathbf{S}_j^*(u)$ has the form $\mathbf{S}_j^*(u) = \mathbf{M}\mathbf{S}_j(u)\mathbf{M}'$.*

Proof: See Appendix A.3.

3.2.3 Coherence and Partial Coherence within the MvLSW setting

We now introduce a measure of cross-dependence between different channels at a particular scale. We can quantify this dependence by defining the wavelet coherence between channels. For our multivariate series we will define the coherence in terms of the wavelet *coherence matrix*.

Definition 3.4 *For scale, j, rescaled time point, $u \in (0,1)$, the wavelet coherence matrix, $\boldsymbol{\rho}_j(u)$ is defined as,*

$$\boldsymbol{\rho}_j(u) = \mathbf{D}_j(u)\mathbf{S}_j(u)\mathbf{D}_j(u). \tag{3.5}$$

Here $\mathbf{S}_j(u)$ is the LWS matrix defined previously. We also define $\mathbf{D}_j(u)$ to be a diagonal matrix whose elements are $S_j^{(p,p)}(u)^{(-1/2)}$.

The (p, q) element of the wavelet coherence matrix, $\rho_j^{(p,q)}(u)$, is the coherence between channels p and q of the series. This individual element can also be expressed as,

$$\rho_j^{(p,q)}(u) = \frac{S_j^{(p,q)}(u)}{\sqrt{S_j^{(p,p)}(u) S_j^{(q,q)}(u)}}. \tag{3.6}$$

Remark. Given this expression it is clear that the coherence between channels will take a value between -1 and 1 at any given point in time. A value close to ± 1 indicates a strong positive/negative linear dependence between channels at that time and scale. A value close to 0 shows there is little or no linear dependence between channels. Setting $p = q$ in Equation (3.6) demonstrates that the diagonal elements of $\boldsymbol{\rho}_j(u)$ are equal to 1. In Fourier analysis a quantity with these properties would generally be referred to as coherency however we will follow the terminology of Sanderson et al. (2010) and refer to it as coherence.

When analyzing the coherence structure of a multivariate signal it may, superficially, appear that two channels are linked as there is significant coherence between them. However, it may in fact be the case that there is not a direct link between them but they are both linked via a third series (see Figure 3.2). To this end we conclude our modelling framework by introducing the wavelet partial coherence. This provides a measure of the coherence between two channels after removing the effects of all other channels. Partial coherence can again be defined in matrix form using the LWS matrix. The definition of wavelet partial coherence below is analogous to the Fourier domain definition developed in Dahlhaus (2000b).

Definition 3.5 *We define the matrix* $\mathbf{G}_j(u) = \mathbf{S}_j(u)^{-1}$ *and the diagonal matrix* $\mathbf{H}_j(u)$ *with elements* $G_j^{(p,p)}(u)^{-(1/2)}$. *The wavelet partial coherence matrix at scale,* j, *and rescaled time,* u, *is defined to be*

$$\mathbf{\Gamma}_j(u) = -\mathbf{H}_j(u)\mathbf{G}_j(u)\mathbf{H}_j(u). \tag{3.7}$$

The off diagonal terms of this matrix are the partial coherences between channels. That is the coherence between the channels after the linear effects of all other channels have been removed.

3.3 Estimation of the MvLSW Spectral Dependence Quantities

In this section we turn our attention to estimating the spectral quantities of a MvLSW signal. Specifically we first consider the estimation of the LWS matrix before turning to the estimation of the wavelet coherence and partial coherence which were introduced in Section 3.2.

First, we define the *empirical wavelet coefficient vector*, $\mathbf{d}_{j,k} = [d_{j,k}^{(1)} \ldots, d_{j,k}^{(P)}]'$ whose elements are the empirical wavelet coefficients for each signal channel

$$\mathbf{d}_{j,k} = \sum_{t=0}^{T-1} \mathbf{X}_t \psi_{jk}(t). \tag{3.8}$$

We use the empirical wavelet coefficient vector to produce the raw *wavelet periodogram matrix*, $\mathbf{I}_{j,k}$:

$$\mathbf{I}_{j,k} = \mathbf{d}_{j,k}\mathbf{d}_{j,k}'. \tag{3.9}$$

Moreover, we denote $I_{j,k}^{(p,q)}$ to be the (p,q)-th entry of the periodogram matrix where $p, q \in \{1, \ldots, P\}$. The raw wavelet periodogram matrix is the starting point for estimating the LWS matrix. In order to achieve a final estimator with the correct properties we explore the asymptotic properties of the raw periodogram matrix as an estimator for this quantity. In particular, given the results in the one-dimensional setting, it is natural to enquire whether the raw wavelet periodogram is biased.

Proposition 3.4 *Let $\{\mathbf{X}_t\}$ be a MvLSW signal with underlying LWS matrix, $\mathbf{S}_j(u)$, and empirical wavelet coefficients, $\{\mathbf{d}_{j,k}\}$. Then*

$$\begin{aligned} E\left[\mathbf{I}_{j,k}\right] &= \sum_{l=1}^{J} A_{jl}\mathbf{S}_l(k/T) + \mathcal{O}(T^{-1}) \quad \textit{and} \\ \mathrm{Var}\left\{I_{j,k}^{(p,q)}\right\} &= \sum_{l=1}^{J} A_{jl}S_l^{(p,p)}(k/T) \sum_{l=1}^{J} A_{jl}S_l^{(q,q)}(k/T) \\ &\quad + \left(\sum_{l=1}^{J} A_{jl}S_l^{(p,q)}(k/T)\right)^2 + \mathcal{O}(2^{2j}/T), \end{aligned}$$

where $A_{jl} =< \Psi_j, \Psi_l >= \sum_\tau \Psi_j(\tau)\Psi_l(\tau)$ for $j, l \in \mathbb{N}$ is the inner product matrix of discrete autocorrelation wavelets (see Nason et al. (2000) or Eckley and Nason (2005) for further details).

Proof: See Appendix A.4.

As in the univariate setting of Nason et al. (2000), the above result establishes that the raw wavelet periodogram matrix is both asymptotically biased and inconsistent. The bias has a particular form consisting of entries in the inner product matrix $\mathbf{A}$. In Cardinali and Nason (2010), the inner product matrix $\mathbf{A}$ is established to be invertible for all Daubechies' compactly supported wavelets. Consequently, the bias of the raw wavelet periodogram matrix estimator in Proposition 3.4 can be corrected. However,

this would still be an inconsistent estimator. Thus, our proposal is to first apply a smoother on the raw wavelet periodogram matrix and then correct the bias. In particular, we use a rectangular kernel smoother with window of length $2M+1$ to produce the smoothed estimator,

$$\tilde{\mathbf{I}}_{j,k} = \frac{1}{2M+1} \sum_{m=-M}^{M} \mathbf{I}_{j,k+m}. \tag{3.10}$$

With such an estimator we establish the following result.

Proposition 3.5 *Assume that* $\sup_{z\in[0,1]} |\sum_{\tau} c(z,\tau)| \leq \infty$. *Then*

$$\begin{aligned} E\left[\tilde{I}_{j,k}^{(p,q)}\right] =& \sum_{l=1}^{J} A_{jl} S_l^{(p,q)}(k/T) + \mathcal{O}(MT^{-1}) + \mathcal{O}(T^{-1}) \\ \mathrm{Var}\left\{\tilde{I}_{j,k}^{(p,q)}\right\} =& \mathcal{O}(2^{2j}/M) + \mathcal{O}(2^{2j}/T). \end{aligned}$$

Proof: See Appendix A.5.

Remark. In the limit, as $T, M \to \infty$, $\mathrm{Var}\left\{\tilde{I}_{j,k}^{(p,q)}\right\} \to 0$. Here, one observes the usual bias-variance trade-off: increasing M reduces the variance but also increases the bias. Moreover, with the additional condition that $M/T \to 0$, then $\left| E\left[\tilde{I}_{j,k}^{(p,q)}\right] - E\left[I_{j,k}^{(p,q)}\right] \right| \to 0$. Thus, one can correct the bias of the smoothed periodogram using the inverse of the inner product matrix $\mathbf{A^{-1}}$. The final smoothed bias-corrected estimator of the LWS matrix is then given by

$$\widehat{\mathbf{S}}_{j,k} = \sum_{l=1}^{J} A_{jl}^{-1} \tilde{\mathbf{I}}_{l,k}. \tag{3.11}$$

We will use the quantity $\widehat{\mathbf{S}}_{j,k}$ to estimate the wavelet coherence and partial coherence. Denote the (p,q)-th entry of $\widehat{\mathbf{S}}_{j,k}$ to be $\widehat{S}_{j,k}^{(p,q)}$ and let $\widehat{\mathbf{D}}_{j,k;T}$ be a diagonal matrix whose elements are $(\widehat{S}_{j,k}^{(p,p)})^{-(1/2)}$. Then, we define the estimator of the wavelet

coherence matrix to be,

$$\widehat{\boldsymbol{\rho}}_{j,k} = \widehat{\mathbf{D}}_{j,k}\widehat{\mathbf{S}}_{j,k}\widehat{\mathbf{D}}_{j,k} \text{ for } j \in \{1,\ldots,J\}, k \in \{0,\ldots,T-1\}. \tag{3.12}$$

The (p,q)-th element of $\widehat{\boldsymbol{\rho}}_{j,k}$ is the estimated time-varying wavelet coherence between channels p and q at level j. Next, define $\widehat{\mathbf{G}}_{j,k} = (\widehat{\mathbf{S}}_{j,k})^{-1}$ and let $\widehat{\mathbf{H}}_{j,k}$ be a diagonal matrix whose elements are $(\widehat{G}_{j,k}^{(p,p)})^{-(1/2)}$. Then, the estimator of the wavelet partial coherence matrix is defined to be,

$$\widehat{\boldsymbol{\Gamma}}_{j,k} = -\widehat{\mathbf{H}}_{j,k}\widehat{\mathbf{G}}_{j,k}\widehat{\mathbf{H}}_{j,k} \text{ for } j \in \{1,\ldots,J\}, k \in \{0,\ldots,T-1\}. \tag{3.13}$$

Thus, the (p,q)-th element of $\widehat{\boldsymbol{\Gamma}}_{j,k}$ is the estimated wavelet partial coherence between channels p and q. Note that the linear dependence of channels p and q on all the other channels are removed in the calculation of wavelet partial coherence. Finally we note that using Slutsky's theorem Slutsky (1925) it follows immediately that $\widehat{\boldsymbol{\rho}}_{j,k}$ and $\widehat{\boldsymbol{\Gamma}}_{j,k}$ are asymptotically unbiased and consistent estimators of the true wavelet coherence matrix and wavelet partial coherence matrix, respectively.

3.4 Applications of the Multivariate LSW model

To illustrate our proposed multivariate locally stationary wavelet process (MvLSW) we now consider two examples. Section 3.4.1 considers a simulated example whilst Section 3.4.2 presents an analysis of multivariate EEG data recorded during a visual-motor experiment.

3.4.1 Simulated Example

We simulate signals using a tri-variate model of the following form, $\mathbf{X}_t = \mathbf{A}_1\mathbf{X}_{t-1} + \mathbf{A}_2\mathbf{X}_{t-2} + \boldsymbol{\xi}_t$, where $\mathbf{A}_1 = 1.51\mathbf{I}_3$, $\mathbf{A}_2 = -0.83\mathbf{I}_3$ and $\boldsymbol{\xi}_t = [\xi_t^1 \ \xi_t^2 \ \xi_t^3]' \sim N(0, \boldsymbol{\Sigma}_t)$. Here $\boldsymbol{\Sigma}_t$ varies across time so that the cross-correlation structure changes from one time region to another. The channels of the series will therefore have a time-varying coherence structure which is known and constant over frequency. The structure is such that there is a peak in the spectral power at frequency $3\pi/16$ which corresponds to the mid point of wavelet level $j = 3$. We simulated 100 tri-variate signals from this model. Using the method proposed in Section 3.3 we estimate the coherence and partial coherence matrices for each simulated signal. In the results reported the Haar wavelet was used in the analysis, although in other simulations we observed that the choice of wavelet made little practical difference for this example. For comparison we also calculate the coherence using both the SLEX method and the method of Ombao and Van Bellegem (OVB) in Ombao and Van Bellegem (2008). For direct comparisons, we have calculated these coherence values for the band of frequencies corresponding to wavelet level $j = 3$.

Figure 3.3 shows the results of the coherence estimation. In particular we note that of the three estimation methods, the proposed MvLSW coherence estimation scheme produces the most faithful overall estimate of the three. Most notably OVB fails to suitably capture the abrupt change in coherence which occurs within this simulated example. SLEX performs slightly better than OVB in terms of capturing the abrupt changes however it fails to consistently match the peaks and troughs of the

coherence. The exception to this is the coherence between channels 1 and 2, where the spectral structure is constant. Here SLEX and OVB have both performed better then our MvLSW method. This is unsurprising given that for this pair the coherence is stationary. This is because OVB can adaptively choose the size of the window so that it matches any changes, if present, on the true spectral quantity. Similarly, the SLEX method chooses the best basis for representing signals and thus can adaptively select the stationary basis if the signal is indeed stationary. The results of partial

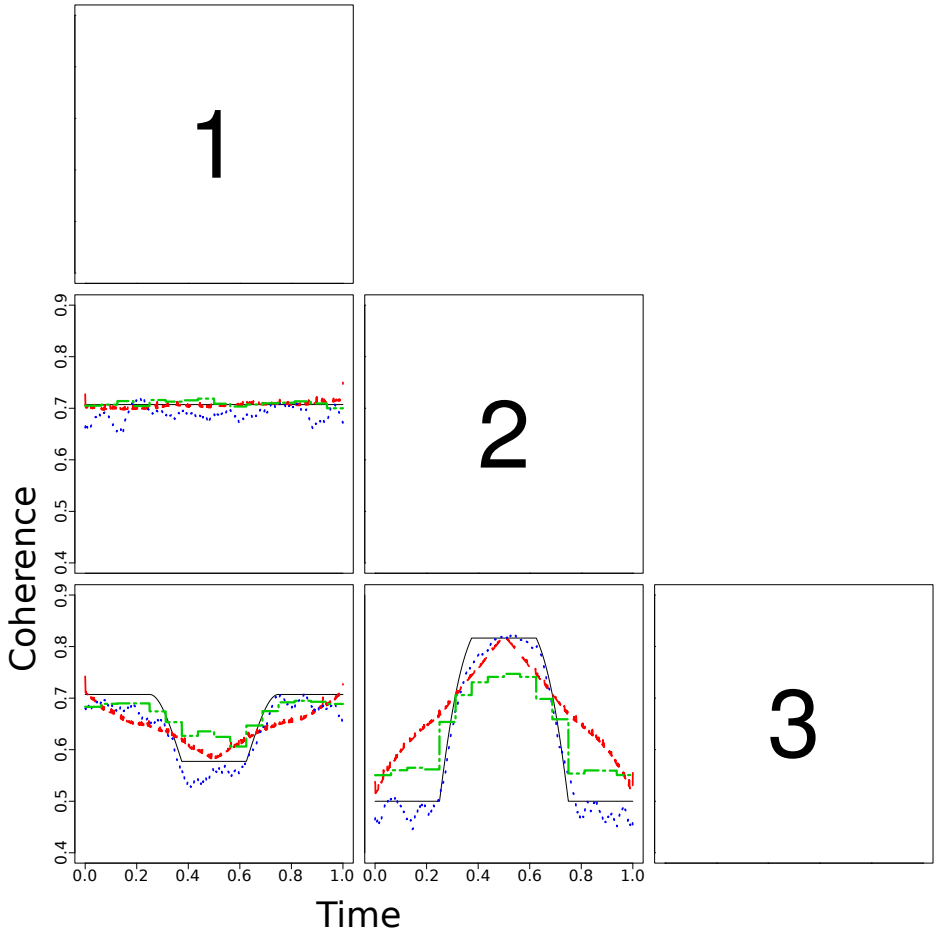


Figure 3.3: Coherence at level $j = 3$: truth (solid) and mean estimate of the coherence obtained from 100 simulations using MvLSW (dotted); SLEX (dotted and dashed) and OVB (dotted).

coherence estimation using the proposed method are shown in Figure 3.4. We draw particular attention to how the wavelet partial coherence estimator is able to capture quite subtle time-localized changes in partial coherence. Comparison of this approach with SLEX and OVB equivalents for partial coherence is left as an avenue for future research, once such methods have been developed in the literature.

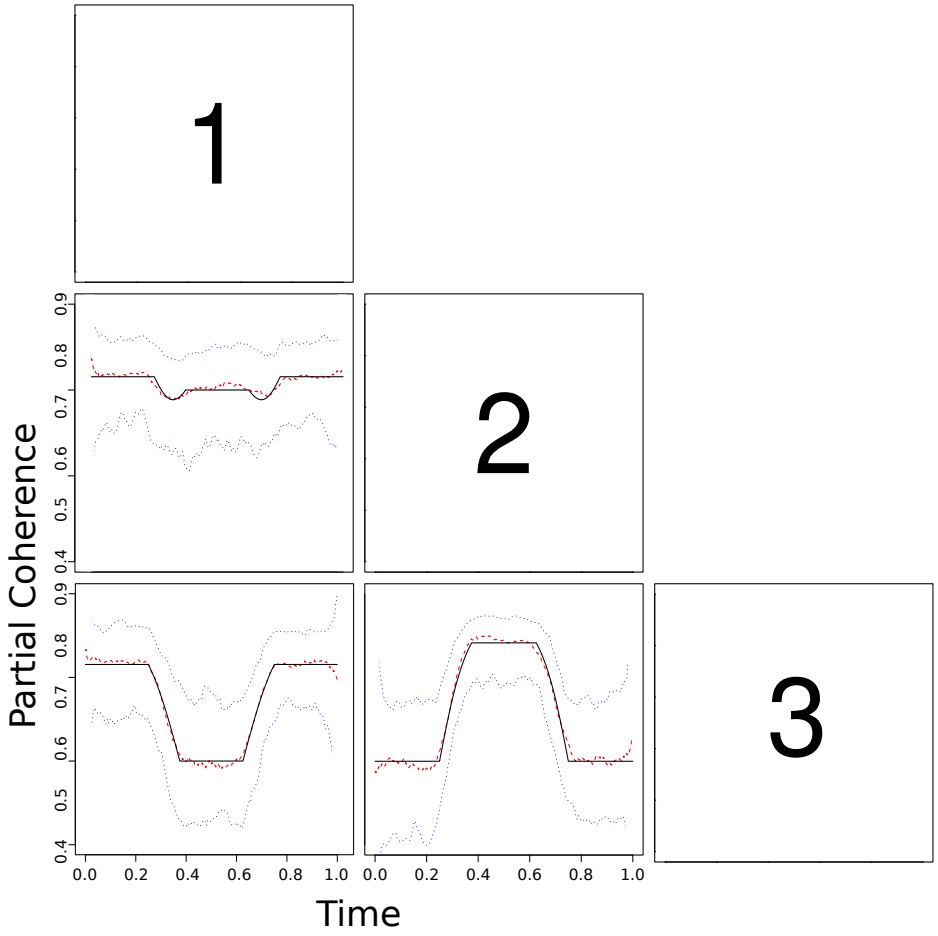


Figure 3.4: Partial coherence at level $j = 3$. Solid lines represent true values, dashed lines represent the mean of 100 simulations and the dotted lines denote approximate 95% point-wise confidence intervals.

3.4.2 EEG Data

Our real data example is a multi-channel electroencephalogram (EEG) recorded from an experiment in which participants are instructed to move a hand held joystick to either the left or right. A 64-channel EEG was recorded at a sampling rate of 512 Hertz and then bandpass filtered at $(0.02, 100)$ Hertz. Each recording epoch was 1000 milliseconds; the instruction (left vs right) was given at time $t = 0$; and the subject responded with a wrist movement between 350 and 450 milliseconds. Here, we selected data for one participant and used 4 channels on the right hemisphere namely FC4 (right fronto-central), FC6 (also right parietal-fronto-central), P4 (right parietal), C4 (right central). This collection is a subset of the channels in Fiecas and Ombao (2011) believed to be engaged in visuo-motor tasks. The positions of these channels are shown in Figure 3.5. Here, we present an analysis of the wavelet spectral quantities computed for level $j = 2$ $(12.5 - 25$ Hertz), which is contained within the conventional beta band. To study the dynamics within each brain region, we estimated the time-varying and level dependent LWS by kernel smoothing the wavelet auto- and cross-periodograms using a smoothing span that was objectively selected by generalized cross-validated gamma deviance criterion developed in Ombao et al. (2001). The Daubechies extremal-phase wavelet 10 vanishing moments was used as the analysing wavelet. We found that by using a smoother wavelet we were able to better capture the dynamics of the coherence and partial coherence of this recording.

We investigated the dynamics of cross-dependence within the brain network by estimating the wavelet coherence and wavelet partial coherence. The point estimates

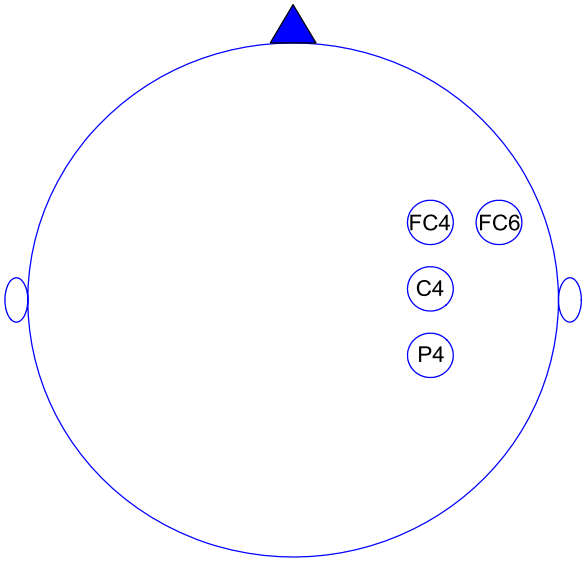
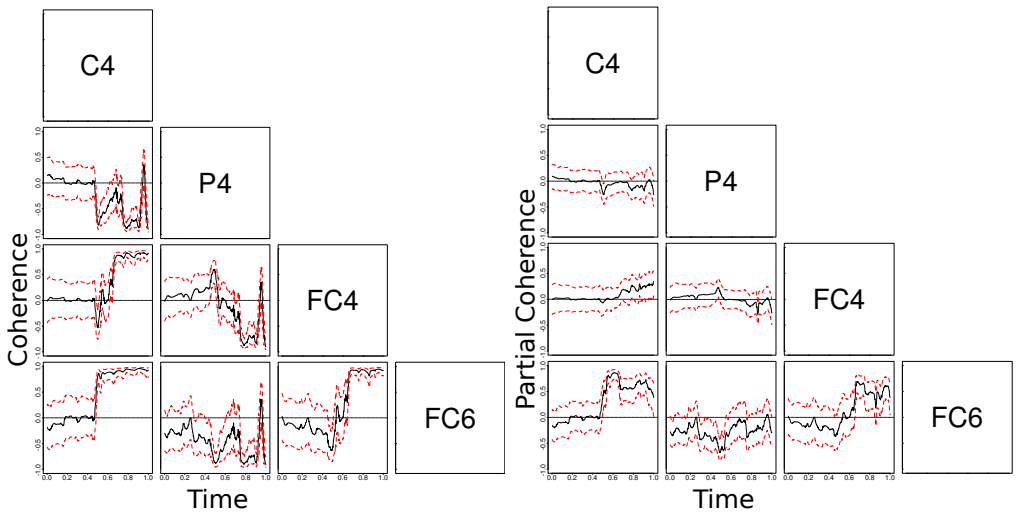


Figure 3.5: Placement of EEG channels included in analysis.

of the wavelet coherence and partial coherence were computed using the quantities in the estimated LWS matrix. The approximate 95% pointwise confidence intervals for coherence and partial coherence were obtained by bootstrap resampling the stochastic component of the MvLSW model. Such an approach was used in Ombao et al. (2000) for inference on the evolutionary SLEX spectrum. Empirical distributions of the Fisher-z transformed wavelet coherence and partial coherence values were constructed based on B bootstrap replicates. Typically one might use $B = 1000$ such replicates. Following ideas from Fourier coherence, see for example Ombao and Van Bellegem (2008), the wavelet coherence and partial coherence estimates were Fisher-z transformed in order to stabilize the variance of the estimator. The scale-shift specific variance of the empirical distribution of the Fisher-z transformed values were extracted and then utilized to compute the approximate 95% pointwise confidence intervals. For ease of interpretation these confidence intervals were then back-transformed to

the scale $(-1, 1)$.



(a) Coherence (b) Partial Coherence

Figure 3.6: Coherence plot (left) and Partial Coherence plot (right) at level $j = 2$. Solid lines represent the estimated values and dashed the approximate 95% point-wise confidence intervals.

The plots displaying confidence bands on the wavelet coherence (see Figure 3.6(a)) suggest that, for the most part, brain activity captured by the P4 channel exhibited no linear dependence with brain activity at the central channels namely C4, FC6 and C4. In contrast, there appears to be a common temporal trend in coherence among the central channels. Early in the signals (immediately following visual instruction) there does not seem to be statistically significant connections. However, at about 400 milliseconds (approximately the time the subject responds to the cue by moving), these central channels become strongly coherent with each other at the beta frequency band. It is interesting to see these brain dynamics during hand movement.

The natural follow-up question is whether or not the links between the central

channels established by the coherence plots are *direct* or *indirect* (i.e., due to a connection0 with some common channel). We addressed this question by using the wavelet partial coherence within the framework of our proposed MvLSW model. In Figure 3.6(b), note that brain activity at FC4 was not directly linked to brain activity at the C4 channel but the link between FC4 and was FC6 was statistically significant beginning at around $t = 400$ milliseconds. Moreover, we observe that there was a statistically significant direct link between FC4 and FC6 – suggesting that the connection between FC4 and C4 observed in the coherence plot was not direct but was in fact related to their common link with the FC6 channel.

The results produced by the proposed MvLSW model are similar to the results from a Fourier-based approach in Fiecas et al. (2010). More importantly, we demonstrate that our proposed model and cross-dependence measure are able to identify an interesting result on the small network of central channels that suggest a direct link between activity at the FC6 channel and each of the FC4 and C4 channels during a visual-motor activity. This finding certainly requires further scientific experiments especially in how these direct connections might be crucial to preserving motor function as well as recovering lost motor function following a major traumatic brain injury. Of course, this analysis is done only on one subject and one will have to develop a more complex model that would take into account brain response variation across many subjects. Nevertheless, the analysis has demonstrated the potential utility and broad impact of the MvLSW model.

3.5 Concluding Remarks

In conclusion, we developed a rigorous, wavelet-based modeling framework which can capture the evolutionary scale-dependent cross-dependence between components of multivariate signals. An associated estimation theory was also established, demonstrating the uniqueness and asymptotic consistency of our spectral estimators. The particular construction which we proposed also permits the identification of time-scale localized coherence and partial coherence. The proposed wavelet partial coherence measure, in particular, can prove useful when considering the linear dependence between a pair of channels as it enables us to decouple the linear effects of other components of the multivariate signal.

Chapter 4

Dynamic Classification of Multivariate Time Series Using the Multivariate Locally Stationary Wavelet Model

Abstract

Methods for the supervised classification of signals generally aim to assign a signal to one class for its entire time span. In this paper we present an alternative formulation for multivariate signals where the class membership is permitted to change over time. Our aim therefore changes from classifying the signal as a whole to classifying the signal at each time point to one of a fixed number of known classes. We assume that each class is characterised by a different stationary generating process, the signal as a whole will however be nonstationary due to class switching. To capture this nonstationarity we use the recently proposed Multivariate Locally Stationary Wavelet model. To account for uncertainty in class membership at each time point our goal is not to assign a definite class membership but rather to calculate the probability of a signal belonging to a particular class. Under this framework we prove some asymptotic consistency results. This method is also shown to perform well when applied to both simulated and accelerometer data. In both cases our method is able to place a high probability on the correct class for the majority of time points.

4.1 Introduction

This paper focuses on a supervised signal classification problem for multivariate signals. The canonical supervised signal classification problem considered within the literature, see for example Kakizawa et al. (1998); Shumway (2003); Huang et al. (2004); Sakiyama and Taniguchi (2004); Caiado et al. (2006); Fryzlewicz and Ombao (2009); Böhm et al. (2010); Liu and Maharaj (2013); Krzemieniewska et al. (2014), may be briefly summarised as follows: Assume that we are given a nonstationary signal of unknown class label, then we seek to assign the *entire* signal to one of N_c different classes, using training data. The implicit assumption within the above, of course, is that the underlying process does not switch between classes.

In practice one can conceive of several situations where such a 'mono-class' assumption might not be appropriate. For example, the nonstationary signal in question might be piecewise (second-order) stationary, with each stationary block representing a particular class structure. To illustrate this we introduce a motivating example using accelerometer data recorded from a movement experiment, one run of which is shown in Figure 4.1. The experiment involves a participant performing a series of activities, namely: walking down a corridor, up a set of stairs and down a set of stairs. The interest in this setting is not to classify the whole signal, but rather to associate a class with each particular activity. As such the inference challenge we address in this article is that of dynamically classifying a nonstationary signal at a given time point into a particular pre-determined class structure.

The problem of classification of signals has a long history dating back to early work

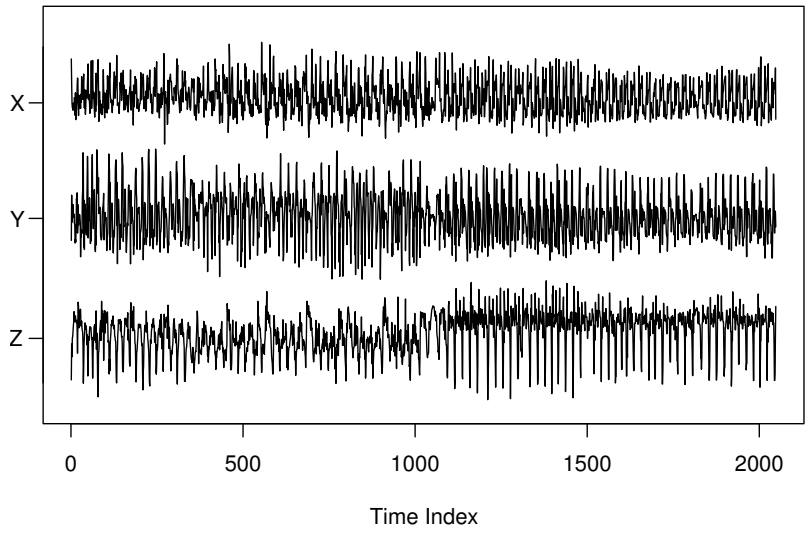


Figure 4.1: Tri-axial accelerometer signal.

on the classification of (second-order) stationary (univariate) signals. For overviews of this area we refer the reader to Shumway (1982). In the nonstationary signal setting one could use various frameworks including nonstationary adaptations of the stationary Fourier basis, see for example Sakiyama and Taniguchi (2004) which adopts the locally stationary Fourier model in Dahlhaus (1997). An alternative Fourier based approach is considered by Huang et al. (2004) and Böhm et al. (2010) who adopt the smooth localised exponentials (SLEX) framework. Of course one need not be restricted to the Fourier basis. For example, Fryzlewicz and Ombao (2009) and Krzemieniewska et al. (2014) use the locally stationary wavelet approach of Nason et al. (2000) for univariate signal classification. In each of these settings the focus is on classifying a signal into one class, i.e. they do not tackle the problems of nonstationarity due to class switching. Thus these approaches are inadequate for classifying

many real systems.

One possible approach to our problem of dynamic classification would be to segment the signal *a priori* and then assign each segment to a particular class. Such an approach is discussed in Krzemieniewska (2013). However such pre-processing can lead to some potential pitfalls. For example, in the case of a high dimensional signal, the differences between classes may be driven by only a small proportion of the channels. This can make segmentation challenging and the overall quality of classification will rely heavily on the segmentation method used. Another possible approach would be to employ a hidden Markov model (HMM). For a review of HMMs we refer the reader to MacDonald and Zucchini (1997) or Cappé et al. (2006). Such an approach is used for classification by Cappé (2002) but is restricted to count data. A HMM framework is also used by Nam et al. (2014) in the related field of changepoint detection. Fitting a HMM has the drawback of being computationally intensive. It also requires the assumption that class transitions are Markovian. In other words the probability of transitioning from one class to another would not depend on time or previous class memberships. In the absence of prior information to support these assumptions such an approach would be difficult to justify. With this in mind we introduce a novel and computationally efficient wavelet based method for classifying a multivariate signal. Our approach estimates the probability of the signal belonging to a particular class at each time point. Importantly our approach, which requires an assumption of local stationarity, does not require any pre-processing of the data.

The method which we introduce is based on the Multivariate Locally Stationary Wavelet model introduced by Park et al. (2014). The Multivariate Locally Stationary

Wavelet model is able to account for changes in both the second order properties of the individual channels of a multivariate signal as well as the linear relationships between channels. For our classification model the nonstationarity in the signal is due to class switching causing the underlying process to change. In this article we focus on the dependence *between* channels by using wavelet coherence. Wavelet coherence has the useful property of being normalised with respect to the local spectral structure. Other methods, such as Huang et al. (2004) or Fryzlewicz and Ombao (2009), normalise the spectral estimates using the global variance of the signal. In our setting, where class membership is a local rather than global, characteristic we must us a local normalisation. Our ultimate goal for classification is to identify the probability of the test signal belonging to each of the classes at a particular time given the observed data. Calculating these probabilities, as opposed to assigning whichever class is closest according some distance measure, will demonstrate the uncertainty in classification.

The remainder of the paper is organised as follows. Section 4.2 provides an overview of the Multivariate Locally Stationary Wavelet model as well as the parameter estimation method which will be used. The main contribution of this paper is contained in Section 4.3 which gives details of our classification method and how it can be applied in practice. Section 4.4 contains two different examples of our method applied to simulated data while Section 4.5 contains an example of our method applied to accelerometer data.

4.2 The Multivariate Locally Stationary Wavelet Model

We now introduce the modelling framework which will be used as the foundation of our classification model, the Multivariate Locally Stationary Wavelet model of Park et al. (2014). This is a multivariate generalisation of the univariate LSW model of Nason et al. (2000). Following Park et al. (2014) let $\mathbf{X}_t = [X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(P)}]'$ be a P-dimensional Multivariate Locally Stationary Wavelet process of length T where $T = 2^J$ for some $J \in \mathbb{N}$. Also let $\mathbf{V}_j(k/T)$ be a lower triangular matrix of functions known as the transfer function matrix and $\{\mathbf{z}_{jk}\}$ be a set of independent random vectors with the properties $E[\mathbf{z}_{jk}] = \mathbf{0}$ and $\text{Var}\{\mathbf{z}_{jk}\} = \mathbf{1}$. Finally let $\{\psi_{j,k}\}$ be the set of discrete wavelet coefficients. $\mathbf{X}_t$ can then be represented in the mean-squared sense as follows,

$$\mathbf{X}_t = \sum_{j=1}^{\infty} \sum_{k} \mathbf{V}_j(k/T)\psi_{j,t-k}\mathbf{z}_{j,k}. \tag{4.1}$$

The transfer function matrix dictates both the auto- and cross-covariance properties of the signal. These properties can be uniquely represented by the Local Wavelet Spectral, LWS, matrix which is defined at scale j and rescaled time point $u = t/T$ as, $\mathbf{S}_j(u) = \mathbf{V}_j(u)\mathbf{V}_j'(u)$. The diagonal elements of the LWS determine the auto-covariance structure of the individual channels of the signal, whilst the off diagonal terms determine the cross-covariance structure between pairs of channels.

Following Park et al. (2014), we define the wavelet coherence at scale j to be the

matrix, $\boldsymbol{\rho}_j(u)$, which has the form,

$$\boldsymbol{\rho}_j(u) = \mathbf{D}_j(u)\mathbf{S}_j(u)\mathbf{D}_j(u), \tag{4.2}$$

where $\mathbf{D}_j(u)$ is a diagonal matrix whose elements are $S_j^{(p,p)}(u)^{(-1/2)}$. The (p,q)-th element of the coherence matrix, $\rho_j^{(p,q)}(u)$, quantifies the strength of any linear relationship between channels p and q at scale j and rescaled time point u and takes a value on the interval [-1,1]. A value close to 1 indicates a strong linear relationship whereas a value close to -1 indicates a strong negative relationship.

To estimate the LWS and coherence matrices of a process we introduce the empirical wavelet coefficient vector at scale j and location k, $\mathbf{d}_{jk} = \sum_t \mathbf{X}_t \psi_{jk}$. This vector can be used to define the raw wavelet periodogram matrix, $\mathbf{I}_{jk} = \mathbf{d}_{jk}\mathbf{d}'_{jk}$. Park et al. (2014) establish that this is a biased and inconsistent estimator of the true LWS matrix, $\mathbf{S}_{jk}$. Consistency can be achieved by smoothing the estimate over time using a rectangular kernel smoother with window size $(2M+1)$. Moreover the bias can be removed using the autocorrelation wavelet inner product matrix, $\mathbf{A}$, with elements $A_{jl} = \sum_\tau \Psi_j(\tau)\Psi_l(\tau)$ where $\Psi_j(\tau) = \sum_k \psi_{jk}(0)\psi_{jk}(\tau)$ (see Nason et al. (2000) or Eckley and Nason (2005) for further details). Hence our estimate of the LWS matrix is given by $\widehat{\mathbf{S}}_{jk} = (2M+1)^{-1} \sum_{m=k-M}^{k+M} \sum_l A_{jl}^{-1} \mathbf{I}_{lm}$. The coherence matrix can then be estimated by substituting $\widehat{\mathbf{S}}_{jk}$ into equation (4.2). In Section 4.3 we will make use of wavelet coherence in order to classify a signal.

4.3 Dynamic Classification

We now consider the classification problem for a Multivariate Locally Stationary Wavelet signal, $\mathbf{X}_t$. The setting which we consider is the following: Assume that at any time, t, $\mathbf{X}_t$ will belong to one of $N_c \geq 2$ different classes where N_c is known. The class membership of $\mathbf{X}_t$ at time t is denoted by $C_X(t) \in \{1, 2, \ldots, N_c\}$. We do not assume that the class membership of $\mathbf{X}_t$ is constant for all time points, nor do we assume that the time spent in a particular class is fixed. Instead we assume that whilst a signal is in a given class it is second order stationary. In other words if $C_X(t) = c,\ \forall t \in \{\tau_1, \ldots, \tau_2\}$, the transfer function matrix, $\mathbf{V}_j(t)$ is a constant, i.e. $\mathbf{V}_j(t) = \mathbf{V}_j^{(c)},\ \forall t \in \{\tau_1, \ldots, \tau_2\}$. The matrix $\mathbf{V}_j^{(c)}$ is the class specific transfer function which has the same lower triangular form as the transfer function matrix described in Section 4.2, however $\mathbf{V}_j^{(c)}$ is constrained to be constant over time. In effect this particular assumed representation means that we can re-express the representation in equation (4.1) as follows. Let $\mathbb{I}_{\{c\}}[C_X(t)]$ be an indicator function which is equal to 1 if $C_X(t) = c$ and 0 otherwise. Then $\mathbf{X}_t$ can be expressed as,

$$\mathbf{X}_t = \sum_k \sum_j \sum_{c=1}^{N_c} \mathbb{I}_{\{c\}}[C_X(k)] \mathbf{V}_j^{(c)} \psi_{jk}(t) \mathbf{z}_{jk}.$$

In effect what we have done here is to re-write the time varying transfer function matrix in terms of constant segments, $\mathbf{V}_j(k/T) = \sum_{c=1}^{N_c} \mathbb{I}_{\{c\}}[C_X(k)] \mathbf{V}_j^{(c)}$.

With this formulation in place it is readily seen that we can also write the LWS of $\mathbf{X}_t$ as,

$$\mathbf{S}_j(u) = \mathbf{V}_j(u) \mathbf{V}_j'(u) = \sum_{c=1}^{N_c} \mathbb{I}_{C_X(u)=c} \mathbf{S}_j^{(c)},$$

where $\mathbf{S}_j^{(c)}$ is the class specific LWS defined as $\mathbf{S}_j^{(c)} = \mathbf{V}_j^{(c)}\mathbf{V}_j^{(c)'}$. Equivalently we can express the time varying coherence matrix of $\mathbf{X}_t$ as, $\boldsymbol{\rho}_j(u) = \mathbb{I}_{C_X(u)=c}\boldsymbol{\rho}_j^{(c)}$.

In the next section we will use the coherence matrix to determine which class the signal belongs to at a particular time. In order to do this we assume that each class has a different coherence matrix, or more precisely, $\exists j$ such that $\boldsymbol{\rho}_j^{(c_1)} - \boldsymbol{\rho}_j^{(c_2)} \neq \mathbf{0}, \forall c_1, c_2 \in \{1, 2, \ldots, N_c\}, c_1 \neq c_2$.

4.3.1 Training Data

To estimate the probability of signal $\mathbf{X}_t$ being in a particular class at a particular time we make use of a set of N_i labelled training signals, the i-th element of which is denoted, $\{\mathbf{Y}_t^{(i)}\}_{i\in\{1,2,\ldots,N_i\}}$. Each of the labelled signals are assumed to have a representation of the form described in Section 4.3. Each training signal will have an associated class function $C_{Y^{(i)}}(t)$ which is known. We estimate the LWS matrix, $\widehat{\mathbf{S}}_{jk;Y^{(i)}}$, for each training signal followed by the coherence matrix, $\widehat{\boldsymbol{\rho}}_{jk;Y^{(i)}}$. This is done using the method described in Section 4.2.

Our ultimate goal for classification is to calculate the probability of the signal belonging to a particular class at a particular time point. To do this we must calculate the likelihood and therefore make distributional assumptions about the estimated coherence. We find in practice that the coherence does not tend to readily fit any standard distribution. We therefore take a Fisher's-z transform of the coherence, the estimates of which are well approximated by a Gaussian distribution, see Fisher

(1915). The transformed coherence for class c, $\boldsymbol{\zeta}_j^{(c)}$ is,

$$\boldsymbol{\zeta}_j^{(c)} = \tanh^{-1} \boldsymbol{\rho}_j^{(c)}. \tag{4.3}$$

The mean of the transformed coherence estimate for class c is thus estimated by averaging the elements of the transformed coherence estimate, $\widehat{\boldsymbol{\zeta}}_{jk;Y_i} = \tanh^{-1} \widehat{\boldsymbol{\rho}}_{jk;Y_i}$, for which $C_{Y^{(i)}}(k) = c$,

$$\widehat{\boldsymbol{\zeta}}_j^{(c)} = \frac{1}{\sum_{i=1}^{N_i} \#(C_{Y^{(i)}}(k/T_i) = c)} \sum_{i=1}^{N_i} \sum_{k \in C_{Y^{(i)}}(k/T_i)=c} \widehat{\boldsymbol{\zeta}}_{kj;Y^{(i)}}. \tag{4.4}$$

In a similar way the variance can also be estimated from the training data.

4.3.2 Selection of Highly Discriminative Coefficients

Following Fryzlewicz and Ombao (2009); Krzemieniewska et al. (2014) we will not use the whole set of transformed coherence coefficients for classification. Instead we use a subset of coefficients which show a significant difference between classes. Using a subset of highly discriminative coefficients will reduce the error in the class probability estimate and also reduce the computational complexity of calculating the log-likelihood. We denote such a subset, which contains the scale and channel indices (j, p, q) for $p < q$, as $\mathcal{M}$. In order to select the appropriate coefficients we rank them according to the distance measure, $\Delta_{jk}^{(p,q)}$, defined as,

$$\Delta_j^{(p,q)} = \sum_{c=1}^{N_c} \sum_{g=c+1}^{N_c} \left| \frac{\widehat{\zeta}_j^{(p,q)(c)} - \widehat{\zeta}_j^{(p,q)(g)}}{\sqrt{\mathrm{Var}\left\{\widehat{\zeta}_j^{(p,q)(c)}\right\} + \mathrm{Var}\left\{\widehat{\zeta}_j^{(p,q)(g)}\right\}}} \right|. \tag{4.5}$$

This distance measure is adapted from the distance measure in Krzemieniewska et al. (2014) and incorporates the variance of the transformed coherence estimates which

can be found empirically using the training data. We select those coefficients which are found to have the largest distance measure.

4.3.3 Classification

Our ultimate goal is to estimate the time varying class membership of the signal, $\mathbf{X}_t$. We do this by estimating the probability of the signal belonging to a particular class at a particular time point. We first estimate the transformed coherence for $\mathbf{X}_t$ denoted as $\widehat{\boldsymbol{\zeta}}_{jk;X}$. Given this estimate we can use Bayes' theorem to obtain,

$$\Pr\left[C(k)=c|\,\widehat{\boldsymbol{\zeta}}_{jk;X}\right] \propto \Pr\left[C(k)=c\right]\mathcal{L}\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\right), \tag{4.6}$$

where $\mathcal{L}(\theta|x)$ is the likelihood and $\Pr\left[C(k)=c\right]$ is a prior probability.

Note: In the absence of prior knowledge we assign an equal prior probability of $1/N_c$ to each class.

Due to the use of the Fisher-z transform we can assume that the distribution of the transformed coherence estimator can be approximated by a Gaussian distribution and so $\mathcal{L}(x|\theta)$ is the Gaussian likelihood function with mean vector, $\mu^{(c)}$ and variance covariance matrix, $\Sigma^{(c)}$. The elements of $\mu^{(c)}$ are the elements of $\zeta_j^{(p,q)(c)}\ \forall\ p,q,j\in\mathcal{M}$. We also define $\widehat{\mu}_k$ which contains the elements of $\widehat{\zeta}_{jk;X}^{(p,q)}\ \forall\ p,q,j\in\mathcal{M}$. The density function, up to a constant factor, can then be expressed as follows:

$$\mathcal{L}\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\,\boldsymbol{\zeta}_j(k/T)=\boldsymbol{\zeta}_j^{(c)}\ \forall j\right) \propto \left|\Sigma^{(c)}\right|^{-\frac{1}{2}}\exp-\frac{1}{2}\left\{(\widehat{\mu}_k-\mu^{(c)})'\left(\Sigma^{(c)}\right)^{-1}(\widehat{\mu}_k-\mu^{(c)})\right\}. \tag{4.7}$$

Since the true mean vectors and variance covariance matrices of $\widehat{\boldsymbol{\zeta}}_{jk;X}$ are not known we substitute estimates taken from the training data described in Section 4.3.1. Computational considerations mean that it is easier to calculate the log-likelihood function, $\ell(x\,|\theta) = \log\{\mathcal{L}(x\,|\theta)\}$. These can be easily related to the probabilities using the following

$$\Pr\left[C(k) = c|\,\widehat{\boldsymbol{\zeta}}_{jk;X}\right] = \frac{\exp\left\{\ell\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\boldsymbol{\zeta}_j(k/T) = \boldsymbol{\zeta}_j^{(c)}\ \forall j\right)\right\}}{\sum_{c=1}^{N_c}\exp\left\{\ell\left(\widehat{\boldsymbol{\zeta}}_{jk;X}\,\middle|\boldsymbol{\zeta}_j(k/T) = \boldsymbol{\zeta}_j^{(c)}\ \forall j\right)\right\}}. \tag{4.8}$$

With the above in place we can consider the probability of misclassification. To this end we define a misclassification at a particular time point t as the highest class membership probability being placed on a class other than the true class. In the following propositions we establish the asymptotic probability of misclassifying a signal of length T.

Proposition 4.1 *Let $\Delta(\widehat{\boldsymbol{\mu}}_k)$ be a divergence criterion for a signal with length T. Also let M_T be the smoothing parameter used for spectral estimation. To ensure an asymptotically consistent and unbiased spectral estimate Park et al. (2014) make the assumptions that $M_T \to \infty$ and $M_T/T \to 0$ as $T \to \infty$. We use the divergence criterion to estimate the class membership at time k/T. In practice we place probabilities on the class memberships however in order to establish the asymptotic properties of the method we use the decision rule, $D(\widehat{\boldsymbol{\mu}}_k)$. For the case of two classes the decision rule is defined as,*

$$D(\widehat{\boldsymbol{\mu}}_k) = \begin{cases} 1 & \textit{(estimate } C(k) = 1\textit{) if } \ \Delta(\widehat{\boldsymbol{\mu}}_k) > 0 \\ 2 & \textit{(estimate } C(k) = 2\textit{) if } \ \Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0 \end{cases}.$$

We show that if the true class membership at time k/T is class 1 then the probability that $D(\widehat{\boldsymbol{\mu}}_k) = 2$ will tend to zero asymptotically, in other words,

$$\lim_{T\to\infty} Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1) = 0$$

Proof: See Appendix B.1.

This result can be generalised to the case of $N_c > 2$ by replacing class 2 with whichever class, other than class 1, has the highest likelihood at location k.

We also consider the asymptotic effect of increasing the Euclidean distance between classes on the misclassification probability.

Proposition 4.2 *Again using the divergence criterion, $\Delta(\widehat{\boldsymbol{\mu}}_k)$, and decision rule, $D(\widehat{\boldsymbol{\mu}}_k)$, defined in proposition 4.1 we consider the two class problem and the distance between classes $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$. We show that for fixed T as the distance between classes increases the probability of assigning the incorrect class tends to zero, in other words,*

$$\lim_{|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|\to\infty} Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1) = 0$$

Proof: See Appendix B.2.

Again this result can be generalised to $N_c > 2$ is the same way as proposition 4.1.

4.4 Simulated Examples

In order to demonstrate how our method works in practice we now present a series of simulated data examples.

4.4.1 Example with Class Specific Autocovariance

The example presented in this section includes signals where both the auto- and cross-covariance structures are dependent upon the time varying class membership. We use a piecewise stationary trivariate autoregressive processes of the form,

$$\mathbf{X}_t = \begin{cases} \phi_1^{(1)}\mathbf{X}_{t-1} + \phi_2^{(1)}\mathbf{X}_{t-2} + \boldsymbol{\xi}_t & \text{if} \quad C_X(t) = 1 \\ \phi_1^{(2)}\mathbf{X}_{t-1} + \phi_2^{(2)}\mathbf{X}_{t-2} + \boldsymbol{\xi}_t & \text{if} \quad C_X(t) = 2 \end{cases}.$$

Here $\{\phi_1^{(1)}, \phi_2^{(1)}\} = \{0.8, -0.5\}$ and $\{\phi_1^{(2)}, \phi_2^{(2)}\} = \{0.9, 0\}$ are the class specific AR coefficients. The set of random elements, $\{\boldsymbol{\xi}_t\}$, are taken from a multivariate normal distribution with zero mean and class specific covariances such that,

$$\boldsymbol{\xi}_t \sim \begin{cases} N(\mathbf{0}, \boldsymbol{\Sigma}^{(1)}) & \text{if} \quad C_X(t) = 1 \\ N(\mathbf{0}, \boldsymbol{\Sigma}^{(2)}) & \text{if} \quad C_X(t) = 2 \end{cases},$$

where,

$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 1 & 0.4 & 0.6 \\ 0.4 & 1 & 0 \\ 0.6 & 0 & 1 \end{bmatrix}, \; \boldsymbol{\Sigma}^{(2)} = \begin{bmatrix} 1 & -0.4 & -0.6 \\ -0.4 & 1 & 0 \\ -0.6 & 0 & 1 \end{bmatrix}. \tag{4.9}$$

We simulate a set of 10 training signals using this model. The training signals each have the same class function which is initially in class 1 and then switches to class 2 half way through the time span. In order to test our method we simulate a group of 100 validation signals. The validation signals all have the same class function which is very different to the one used in the training set. This class function is initially in class 1 but switches 7 times at irregularly spaced intervals. We estimate the class membership probabilities for the validation signals using the method outlined in Section 4.3 and then take the mean.

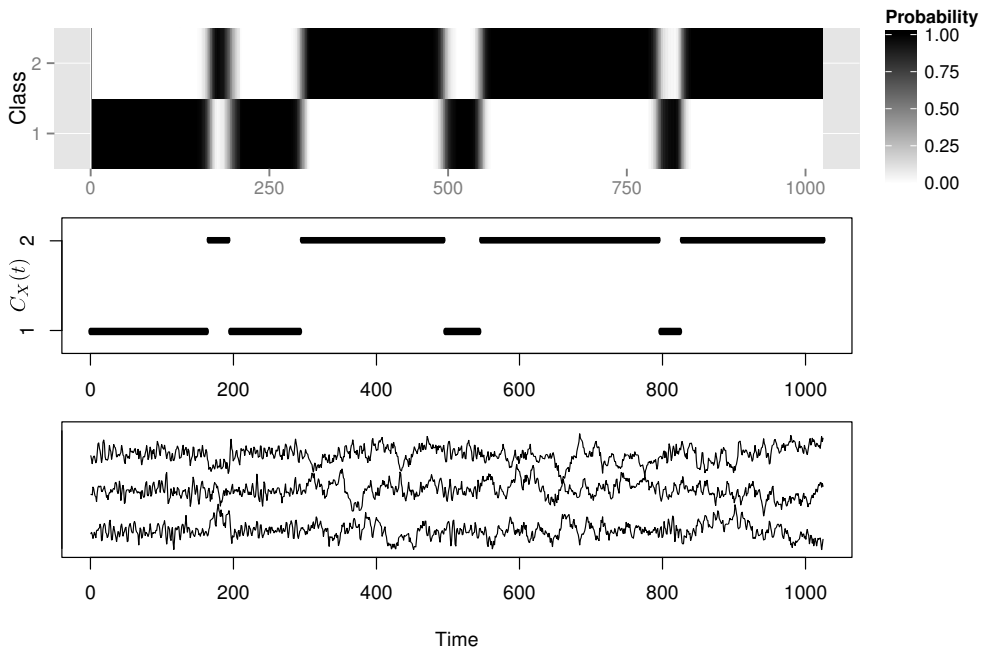


Figure 4.2: The upper plot shows the mean class membership probabilities for the 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time.

The results of this are shown in Figure 4.2. We can see that the mean class probability is consistently high for the true class which shows that our method has performed well in terms of identifying the most likely class for a given time point. We also note that there is a small region of uncertainty around the class transitions which demonstrates that it is more difficult to classify in these regions. Looking at the lower plot in Figure 4.2 we can see that it is possible to identify the class membership visually as the signals autocovariance structure changes noticeably with class due to the changing AR coefficients. In the following sections we will explore examples where this is not the case.

4.4.2 Example with Constant Auto-covariance

We now consider an example with a class specific cross-covariance structure and a constant auto-covariance structure. We again use an autoregressive process where, unlike the previous example, the AR coefficients are not class specific. The general form of the signals is therefore,

$$\mathbf{X}_t = 0.8\mathbf{X}_{t-1} - 0.5\mathbf{X}_{t-2} + \boldsymbol{\xi}_t, \quad \forall t \in \{0, T-1\}. \tag{4.10}$$

The set of random elements, $\{\boldsymbol{\xi}_t\}$ again follow a normal distribution with zero mean and covariances defined in equation (4.9).

Our example is based on a set of 10 training signals and 100 validation signals. The training signals all have the same simple class function as in the previous section, the validation signals all have the same class function which starts in class 2 and switches seven times at irregular intervals. We calculate the class membership probabilities for the validation signals and take the mean, the results are shown in Figure 4.3. Looking at the lower plot in Figure 4.3 we see that for this example it is very challenging to discern the class visually as the auto-covariance structure is constant. The upper plot indicates that despite this our method is still performing to a similar level of accuracy as for the example in Section 4.4.1.

4.4.3 Example with Three Classes

Our final simulated example considers a scenarion where $N_c > 2$. A third class is added to the example in Section 4.4.2. The AR coefficients will remain constant as in equation (4.10) however for time points where $C_X(t) = 3$ the random elements $\{\boldsymbol{\xi}_t\}$

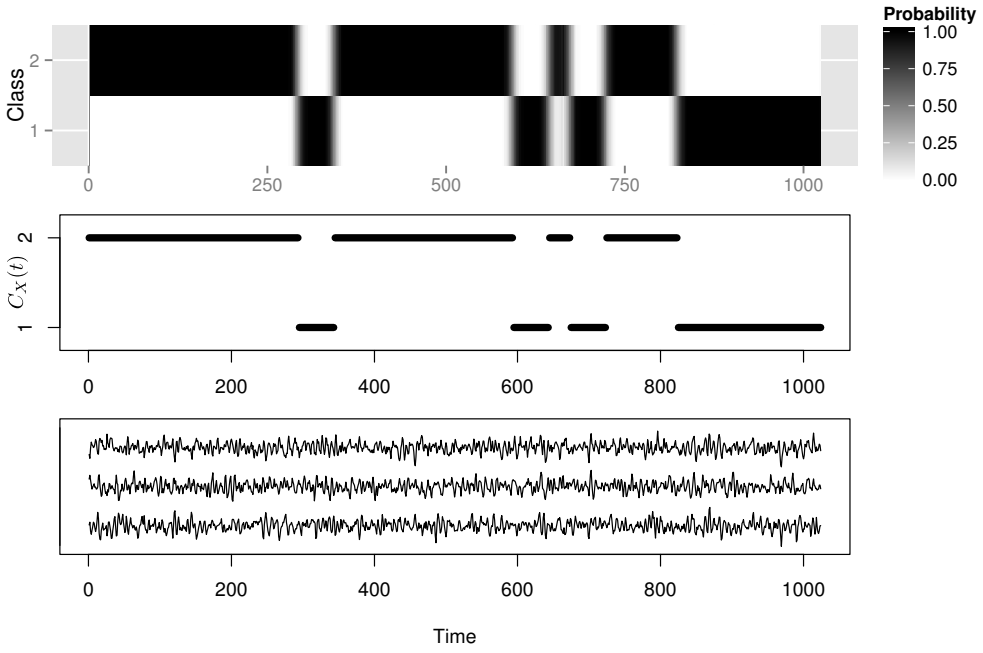


Figure 4.3: The upper plot shows the mean class membership probabilities for the group of 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time

will be taken from a normal distribution with covariance matrix given by,

$$\mathbf{\Sigma}^{(3)} = \begin{bmatrix} 1 & 0.4 & -0.6 \\ 0.4 & 1 & 0 \\ -0.6 & 0 & 1 \end{bmatrix}$$

For this example we again use a set of 10 training signals. Each of the training signals has a class function which cycles through the three classes from 1 to 3 twice. We simulate one group of 100 validation signals which have a class function which also cycles through the class but in reverse order. Figure 4.4 shows the mean class probabilities for the validation signals. Note that our method is able to place a high probability on the correct class for the majority of time points. It is however possible to see

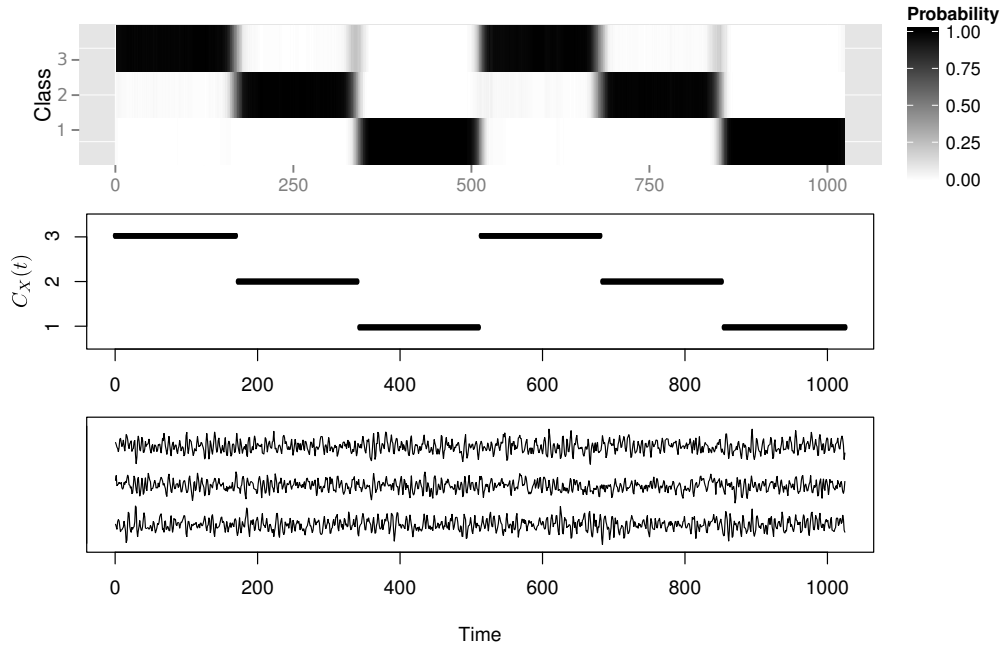


Figure 4.4: The upper plot shows the mean class membership probabilities for the group of 100 validation signals. The lower plot shows one of the validation signals. The middle plot shows the true class membership over time.

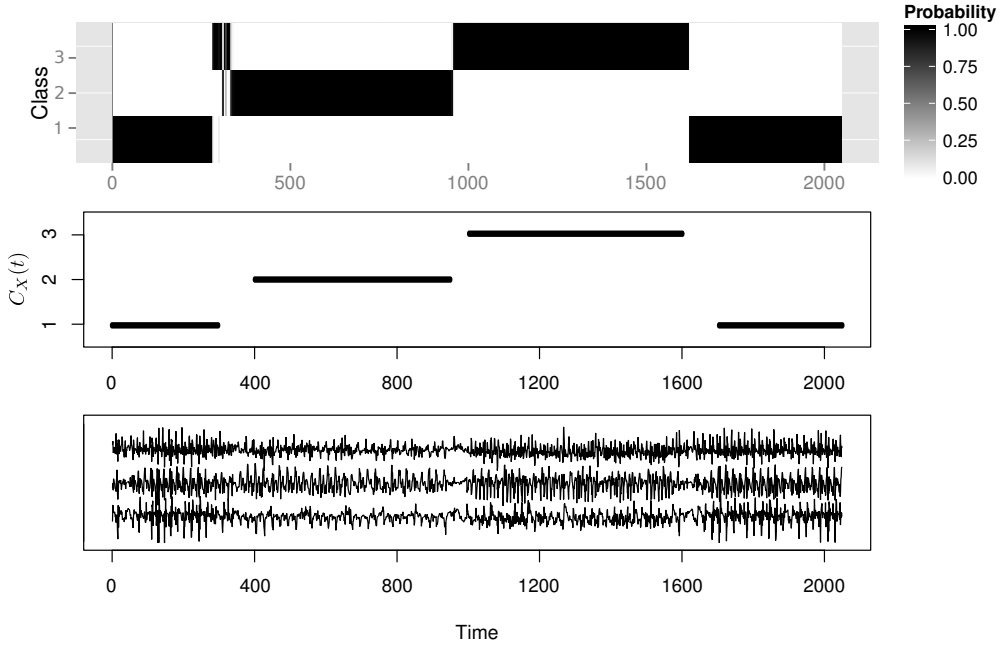
that there are slightly larger regions of uncertainty around the class transitions. This demonstrates that by adding a third class we have made the classification problem more challenging leading to greater uncertainty.

4.5 Accelerometer Data Example

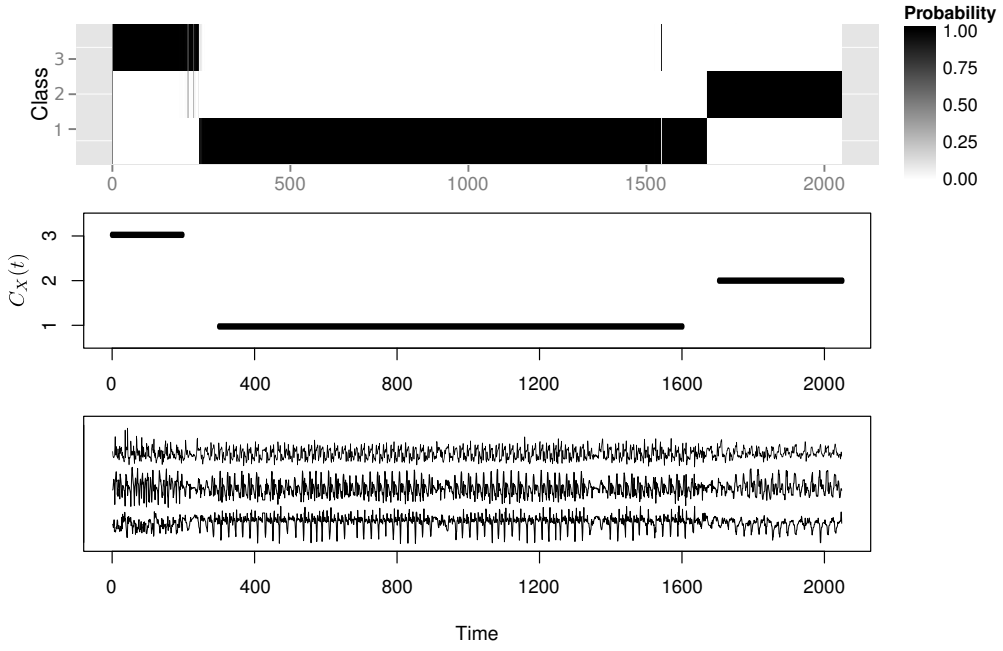
Finally we turn to an example based on tri-axial accelerometer data. A participant is asked to walk normally following a route including a corridor and several flights of stairs whilst wearing a tri-axial accelerometer which has a recording frequency of 20Hz. The experiment is repeated 13 times in total following three different routes, the accelerometer records continuously during each repetition. For 6 of the repetitions the

participant walks along the corridor up the stairs and down the stairs before walking along the corridor again, we will refer to this as Route A. For another 6 repetitions the participant walks down the stairs, along the corridor twice and then up the stairs, we will refer to this as Route B. For the 13th repetitions the participant walks up the stairs, down the stairs and then along the corridor, we refer this as Route C. Each repetition lasts just over 100 seconds and so each recording is trimmed to be of length $T = 2048$.

To illustrate our method we randomly select one repetitions each of Routes A and B and as well as the single repetitions of Route C to classify. The other 10 repetitions will be used as a training set. We adopt a three class model with class 1 being walking along the corridor, class 2 being walking up the stairs and class 3 being walking down the stairs. Figure 4.5 shows the classification results for the signals of Routes A and B. In both cases the true class is given a high probability for nearly all time points, the only exception to this being around the first transition in the Route A signal where the highest probability is placed on class 3 when the true class is either 1 or 2. The middle plots show the true class memberships, it is noticeable that there are very clear shifts in the probabilities which follow the true class memberships.



(a) Route A



(b) Route B

Figure 4.5: Class probabilities for Routes A and B. The upper plots show the estimated class probabilities. The lower plots show the accelerometer recordings. The middle plot shows the true class memberships.

Figure 4.6 shows the results of our classification method performed on the Route C repetition. Since this repetition follows Route C the resulting signal is unlike any in the training data which all follow Route A or Route B. Looking at Figure 4.6 we see that our method is able to place a high probability on the true class for the majority of time points meaning that we can say what activities the participant is performing during Route C.

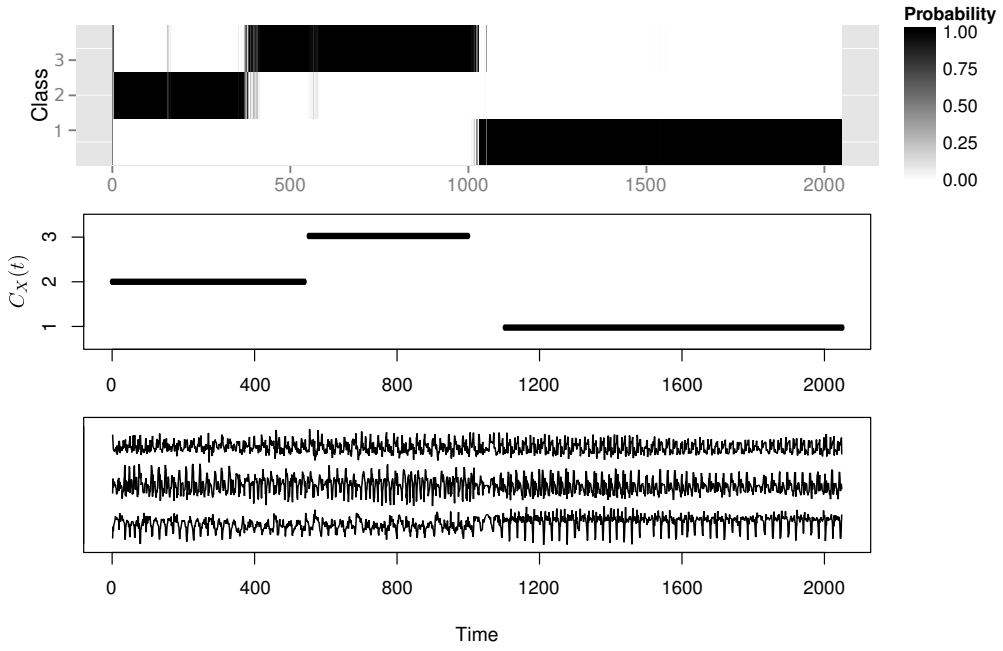


Figure 4.6: Class probabilities for Route C. The upper plots show the estimated class probabilities. The lower plots show the accelerometer recordings. The middle plot shows the true class memberships.

4.6 Conclusion

In this article we proposed a classification method for signals where the class membership is permitted to change over time. Such a model is distinct from the majority

of classification methods which seek to assign a signal to one class for all time points. Our method makes use of a set of labelled training signals to estimate the true spectral properties of each class. Likelihood methods are then used to calculate the probability of the signal being in each class at a particular time point. We also demonstrate this method using both simulated data examples and a real accelerometer data set.

Chapter 5

Wavelet Spectral Confidence Intervals and a Test for Coherence

5.1 Introduction

Recent advances in multi-channel data recording techniques have led to an increased need to analyse multivariate time series. This is particularly true in the medical sector where modelling the dependencies between components of a series can reveal much about the physical processes being studied. In general we cannot assume that the series obtained from these experiments are second order stationary. It is more reasonable to assume that the underlying process exhibits some form of evolving stochastic structure in terms of both the individual components (e.g. autocovariance) *and* the dependence between components. One method for representing such series is the multivariate locally stationary wavelet (MvLSW) model introduced by Park et al. (2014). The MvLSW framework provides a representation of time series with

smoothly changing second order strucutre. Being a wavelet based representation the MvLSW model decomposes the series in terms of frequency bands known as wavelet levels. This is important as knowing which wavelet levels contribute to the variance of a series can reveal much about the generating process. Park et al. (2014) also discuss the modelling of dependencies between components of a multivariate series using wavelet coherence. Wavelet coherence gives a measure of the linear dependence between two components decomposed in terms of time and wavelet level.

Park et al. (2014) establish that MvLSW series can be uniquely represented by the time and level dependent matrix of functions known as the local wavelet spectral (LWS) matrix. They also provide associated estimation theory and prove the asymptotic properties of the estimate.

In this work we further develop this framework by developing confidence intervals for the LWS estimate and a hypothesis test for coherence. The development of confidence intervals permits a better understanding of the uncertaintly of the estimate and can be useful in determining which elemets of the LWS are nonzero and contribute to the structure of the series.

Conversely the development of a hypothesis test enables us to identify which components of the series are dependent. This enables us to build up a more complete dependence structure. Using wavelet coherence for this test allows us to answer the additional questions of which wavelet levels are driving any dependence and for which time points is this dependence present? Such questions are important as the dependencies between components may be localised to a particular level or short span of time points which may be easily missed by a method which is not localised.

Previous work on a test for coherence includes a test in the stationary Fourier setting is discussed in Priestley (1981b), an implementation of a stationary Fourier based test is also described in Shumway and Stoffer (2000). In both cases the test assumes second order stationarity in the series and so will give no information about the time location of any dependence between components.

The rest of this chapter proceeds as follows. Section 5.2 contain the calculation of confidence intervals for the estimated LWS matrix. We demopnstrate the accuracy of our confidence interval by applying it to a known LWS matrix and comparing with confidece intervals obtained by bootstrapping. We also calculate a confidence interval for an LWS estimate obtained from a real EEG recording. In Section 5.3 we turn our attention to defining a hypothesis test for coherence between components. Section 5.4 contains a simulation study demonstrating the effectivness of our method on time series simulated from a range of stationary and nonstationary models. For the stationary models we compare our method to the Fourier based method described in Shumway and Stoffer (2000). Our hypothesis test is also applied to a real EEG recording in Section 5.5. Finally Section 5.6 contains some conclusions and discussion.

5.2 Local Wavelet Spectral Matrix Confidence Intervals

We now consider how one might obtain point-wise confidence for estimates of the LWS matrix, $\widehat{\mathbf{S}}_{j,k}$. Recall from Chapter 3 the LWS estimate is obtained by correcting the smoothed wavelet periodogram, $\widehat{\mathbf{S}}_{j,k} = \sum_l \boldsymbol{A}_{jl}\tilde{\boldsymbol{I}}_{j,k}$ where $\tilde{\boldsymbol{I}}_{j,k} = (2M +$

$1)^{-1}\sum_{m=k-M}^{k+M}\boldsymbol{I}_{j,m}$, and that the raw wavelet periodogram is obtained from the wavelet coefficients such that, $\boldsymbol{I}_{j,k}=\boldsymbol{d}_{jk}\boldsymbol{d}'_{jk}$ and $\boldsymbol{d}_{jk}=\sum_t \boldsymbol{X}_t\psi_{jk}$.

The first step towards calculating this confidence interval is to derive the variance of the (p,q)-th element of the estimated LWS matrix, $\{\sigma_j^{(p,q)}(k/T)\}^2$. Our approach, which we describe below, generalises the result established by Nason (2013) for the univariate setting. The variance can be written in terms of the smoothed periodogram as follows:

$$\left\{\sigma_j^{(p,q)}(k/T)\right\}^2 = \text{Var}\left\{\widehat{S}_{j,k}^{(p,q)}\right\} = \sum_{l_1=1}^{J}\sum_{l_2=1}^{J} A_{jl_1}^{-1}A_{jl_2}^{-1}\,\text{cov}\left(\tilde{I}_{l_1,k}^{(p,q)},\tilde{I}_{l_2,k}^{(p,q)}\right). \tag{5.1}$$

The covariances between elements of the smoothed periodogram can also be written in terms of the raw wavelet periodogram,

$$\text{cov}\left(\tilde{I}_{l_1,k}^{(p,q)},\tilde{I}_{l_2,k}^{(p,q)}\right) = \frac{1}{(2M+1)^2}\sum_{m_1=k-M}^{k+M}\sum_{m_2=k-M}^{k+M}\text{cov}\left(I_{l_1,m_1}^{(p,q)},I_{l_2,m_2}^{(p,q)}\right). \tag{5.2}$$

The final step is to derive the covariance between elements of the raw periodogram. In doing so we first need to introduce a variant of autocorrelation wavelet inner product matrix, A_{jlh}^{λ}. The elements of this are defined as, $A_{jlh}^{\lambda}=\sum_\tau \Psi_{jl}(\lambda+\tau)\Psi_h(\tau)$ with $\Psi_{jl}(\tau)=\sum_k \psi_{j,k}\psi_{l,k+\tau}$ being defined in Fryzlewicz and Nason (2006).

Proposition 5.1 *Let $\mathbf{I}_{j,k}$ be the raw periodogram matrix calculated from a series with true LWS matrix $\mathbf{S}_j(k/T)$, the covariance between the elements of the periodogram matrix is,*

$$\begin{aligned}\text{cov}\left(I_{j,k}^{(p,q)},I_{l,m}^{(p,q)}\right) =& \sum_{h=1}^{J} A_{jlh}^{k-m}S_h^{(p,q)}\left(\frac{k+m}{2T}\right)\sum_{h'=1}^{J}A_{jlh'}^{k-m}S_{h'}^{(p,q)}\left(\frac{k+m}{2T}\right),\\ &+\sum_{h=1}^{J} A_{jlh}^{k-m}S_h^{(p,p)}\left(\frac{k+m}{2T}\right)\sum_{h'=1}^{J}A_{jlh'}^{k-m}S_{h'}^{(q,q)}\left(\frac{k+m}{2T}\right)+\mathcal{O}(T^{-1}).\end{aligned} \tag{5.3}$$

Proof: See Appendix C.1

It was noted by Nason (2013) that, under mild regularity conditions, $\tilde{\boldsymbol{I}}_{j,k}$ and therefore $\widehat{\mathbf{S}}_{j,k}$ is asymptotically normal. This result follows from Schuster (1972). Consequently an (approximate) $100(1-\alpha)\%$ confidence interval for $\widehat{\mathbf{S}}_{j,k}$ can be defined as follows,

$$\left[\widehat{\mathbf{S}}_{j,k} - z_{\alpha/2}\sigma_j^{(p,q)}(k/T), \widehat{\mathbf{S}}_{j,k} + z_{\alpha/2}\sigma_j^{(p,q)}(k/T)\right],$$

where $z_x = \Phi(x)$ the standard normal cumulative distribution function.

Simulated Example: We illustrate the behaviour and accuracy of our analytic LWS confidence intervals by calculating a 95% confidence interval for a known LWS and comparing them to those obtained by parametric bootstrapping as in Park et al. (2014). The known LWS is tri-variate with a series length of $T = 512$ and therefore $J = 9$ levels. The only nonzero power is located in level $j = 2$. For this level some of the LWS elements are time varying thus making the overall process nonstationary.

As in Park et al. (2014) our procedure for calculating the bootstrap confidence intervals is to simulate a series from the true LWS and then estimate the LWS from the simulated series. This is repeated for 1000 simulated series. The 0.025 and 0.975 quantiles of the estimates for each scale and location point is taken to be the 95% confidence interval. The smoothing parameter used for both the variance calculation and bootstrapping is $M = 100$ which we feel is a realistic choice given the smoothness of the true local wavelet spectral. The results are shown in Figure 5.1.

Looking at Figure 5.1 it is clear that the confidence intervals obtained analytically are very close to those obtained by bootstrapping. We note that for all spectral

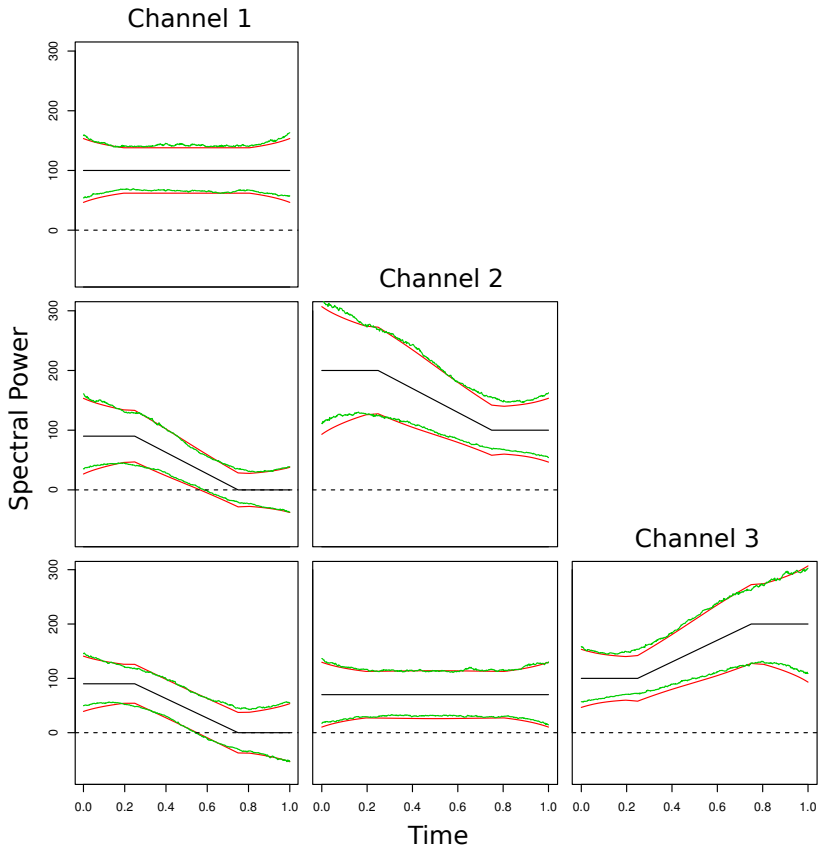


Figure 5.1: Local wavelet spectral matrix confidence intervals. The black lines show the true values, the red lines show the 95% confidence intervals obtained by variance calculation and the green lines show those obtained by bootstrapping. The dotted line indicates zero spectral value.

estimates the confidence intervals become slightly wider close to the start or end of the series. This is not all together surprising given the well known effects whereby the smoothing kernel overlaps with the start or end of the series and so will contain fewer data points. This causes an increase in variability which has been taken into account in our calculation of equation (5.2). Looking closely at the plots we see that some of the largest deviations happen when the trend of the true spectral elements changes. The bootstrap confidence intervals show a smoother transition than those

obtained analytically.

EEG Example: Our real data example again uses the multi-channel electroencephalogram recording used in Section 3.4.2. Recall that a 64-channel recording was taken from an experiment in which participants are instructed to move a hand held joystick to either the left or right. The sampling rate of the recording was 512 Hertz and it was bandpass filtered at (0.02, 100) Hertz. The recording length was 1000 milliseconds; the instruction (left vs right) was given at time $u = 0$; and the subject responded with a wrist movement between 350 and 450 milliseconds. Here, we selected data for one participant who was given the left instruction. In this Chapter we look at 3 channels on the right hemisphere namely FC4 (right fronto-central), FC6 (also right fronto-central), and C4 (right central). The positions of these channels are shown in Figure 5.2. We estimated the LWS matrix using a smoothing span that was objectively selected by generalised cross-validated gamma deviance criterion developed in Ombao et al. (2001). We constructed a 95% confidence interval for the estimate following the procedure from Section 5.2.

Figure 5.3 shows the resulting estimate and confidence interval for level $j = 3$, which corresponds to the frequency range 6.25-12.5Hz. These plots show a drop in the spectral and cross-spectral power for all elements of the local wavelet spectral matrix estimate at this level. The confidence intervals also indicate that for this level the true LWS is likely to be nonzero particularly in the first half of the time series.

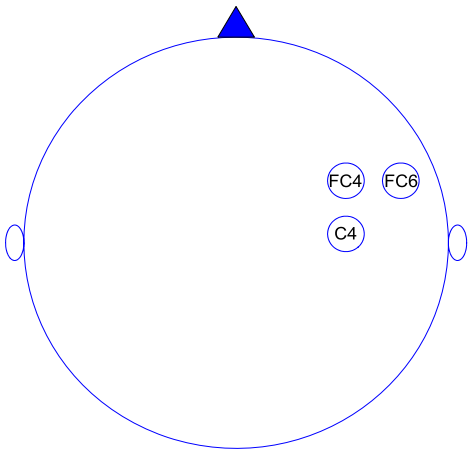


Figure 5.2: Placement of EEG channels included in analysis.

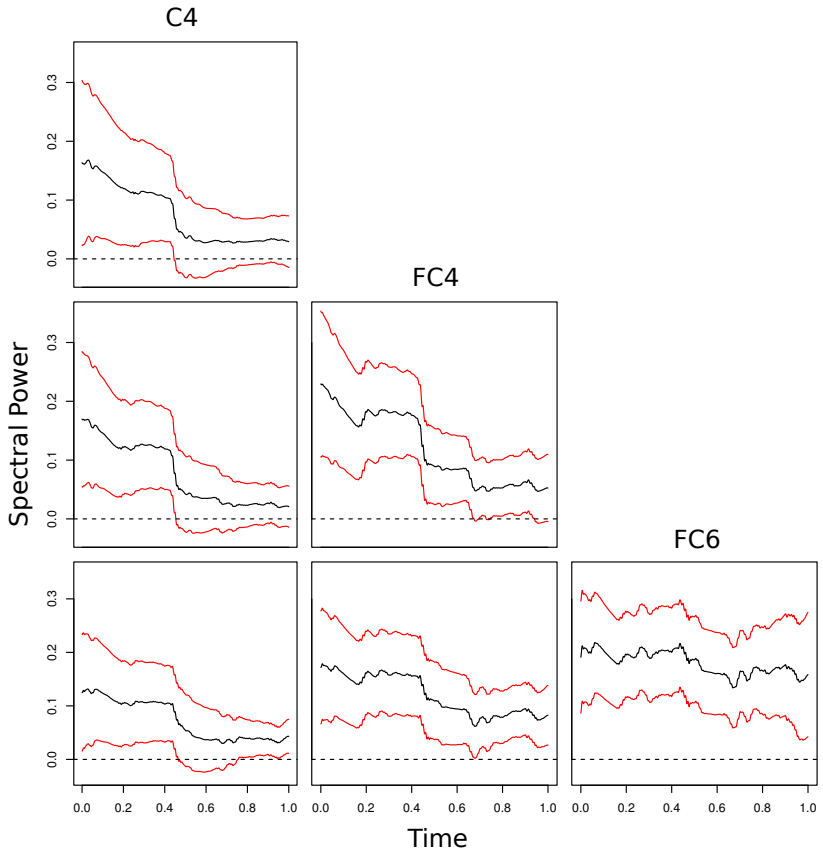


Figure 5.3: EEG spectral estimates for level $j = 3$. The estimate is shown by the black line, the red line show the 95% confidence interval.

5.3 A Test of Coherence

Given the possibility of constructing an (approximate) confidence interval it is natural to also consider the challenge of a test of coherence. This test is performed simultaneously on all channel pairs to highlight which, if any, show statistically significant coherence. We base this test on the MvLSW model which allows us to pinpoint the time and scale locations which are contributing to any coherence. We consider a P dimensional time series $\mathbf{X}_t$ which has a representation under the MvLSW model. Recall from Chapter 3 that for a pair of channels, p and q, at scale j and location k the coherence is defined as, $\rho_j^{(p,q)}(k/T) = S_j^{(p,q)}(k/T)/\sqrt{S_j^{(p,p)}(k/T)S_j^{(q,q)}(k/T)}$. The null and alternative hypotheses for the test of coherence are therefore,

$$H_0 : \rho_j^{(p,q)}(k/T) = 0,$$

$$H_1 : \rho_j^{(p,q)}(k/T) \neq 0.$$

From the definition of the coherence we see that these hypotheses can equivalently be written in terms of the elements of the LWS matrix such that,

$$H_0 : S_j^{(p,q)}(k/T) = 0,$$

$$H_1 : S_j^{(p,q)}(k/T) \neq 0.$$

Since the true LWS is not generally known we make use of the estimate, $\widehat{S}_{j,k}^{(p,q)}$ defined in Chapter 3. As stated in Section 5.2, $\tilde{I}_{j,k}^{(p,q)}$, and therefore $\widehat{S}_{j,k}^{(p,q)}$ is asymptotically normal. The test statistic which we form is thus $z_{j,k}^{(p,q)} = \widehat{S}_{j,k}^{(p,q)}/\widehat{\sigma}_{j,k}^{(p,q)}$. Under the null hypothesis, H_0, the true cross spectrum is zero, $S_j^{(p,q)}(k/T) = 0$, and it follows that the test statistic will have zero mean and unit variance. We can therefore say

that,

$$\Pr\left\{\left|z_{j,k}^{(p,q)}\right| \geq x\right\} \approx 2\left[1-\Phi(x)\right].$$

We apply this test to all channel pairs, $p, q \in \{1, P\},\ p < q$, locations $k \in \{0, T-1\}$ and scales $j \in \{1, J\}$. For a multiple hypothesis test such as this it is necessary to control the rate of false positives. To achieve this we propose using the false discovery rate, FDR, procedure of Benjamini and Hochberg (1995). The simulation study, which is reported below, demonstrates that this method works well. However we do note that this does not account for possible dependencies between the test statistics.

5.4 Simulation Study

In this section we consider the performance of our test of coherence. We consider two different situations in this study, Section 5.4.1 considers stationary time series models while Section 5.4.2 considers nonstationary time series models. In the stationary setting we will compare with the Fourier based test described in Shumway and Stoffer (2000). This method is designed for stationary time series and so is not time localised. We apply the hypothesis test to all available Fourier frequencies and control the false discovery rate using FDR. We will refer to this method as the Fourier test.

The aim of our simulation study in Section 5.4.1 is to compare these two methods in terms of their ability to identify the presence of coherence a pair of components as well as the false discovery rate when the components are in fact independent. Since the Fourier based test does not give a time localised test we count a positive result for any frequency as indicating the pair of components are dependent across all time

points. Conversley our MvLSW based method tests for the presence of coherence across time as well as frequency. Within this setting therefore a positive result at any level is an indication of coherence for that time point only.

5.4.1 Stationary Model Simulations

The first set of simulations which we consider in our simulation study are all second order stationary. That is to say both the autocovariance structures of the individual components *and* the coherences between components do not vary with time. The stationary models which we will use are as follows,

S1: An AR(1) model with parameter equal to -0.9.

S2: An AR(1) model with parameter equal to 0.5.

Under each model we simulate two independent component and then take a linear combination of these to create a bivariate series with correlated components. The values of the correlation, r, which we will test are 0.9, 0.8 and 0.7. Both the Fourier and MvLSW coherence tests are then applied to this bivariate series. The power of the test is calculated as the proportion of time points which yield a positive result. We also estimate the false positive rate by applying both tests to a bivariate series with independent components and recording the proportion of time points which yield a positive test result. We also test each model on a tri-variate series which we denote as **S3**. The channels of this model have the same AR(1) form as model **S1**, however we now simulate three independent components and take linear combinations to achieve correlations between channels. Channels 1 and 2 are constructed to have

a correlation of 0.9, channels 1 and 3 have a correlation of 0.8 and channels 2 and 3 have a correlation of 0.7. The false discovery rate is also tested for this model by applying the tests to three independent components. In all cases we consider 1000 simulated series and perform the test across all time points.

Model	r	Fourier	MvLSW
S1	0.9	1.000	0.996
	0.8	1.000	0.984
	0.7	1.000	0.955
S2	0.9	1.000	0.950
	0.8	1.000	0.860
	0.7	1.000	0.655
S3	$\{0.9, 0.8, 0.7\}$	1.000	0.979

Table 5.1: Simulation Study Results

Model	Fourier	MvLSW
S1	0.055	0.067
S2	0.042	0.001
S3	0.048	0.051

Table 5.2: Simulation Study False Discovery Rate

The results of the power tests are shown in Table 5.1 while the false discovery rate results are shown in Table 5.2. Looking at Table 5.1 we see that, unsurprisingly, the Fourier test has higher power than the MvLSW test in the stationary setting.

In fact it successfully identifies coherence in every simulation. However the MvLSW still performs reasonably well most notably for those models with high correlation, ($r = 0.9$). These results are not unexpected as the Fourier test is able to exploit the stationarity of the simulated models while our MvLSW test must allow for possible nonstationarity. Looking at the false discovery rates in Table 5.2 we see that the Fourier and MvLSW tests perform similarly with false positive rates at around the 5% level.

5.4.2 Nonstationary Model Simulations

The MvLSW coherence test is able to identify the particular time points which contribute to coherence and can thus be used on nonstationary time series. We therefore test this method on some simulated models with nonstationary coherence structures. As the Fourier test is not able to deal with nonstationarity we omit it from this simulation study.

The models which we use may have either stationary or nonstationary autocovariance structures but all will be constructed to have nonstationary coherence strucutres. Specifically the model forms we use are as as follows,

N1: An AR(1) model with parameter equal to -0.5.

N2: An time varying AR(1) model with parameter equal to $0.5 - t/T$.

N3: An MvLSW process with constant spectral power at level $j = 2$.

For each model we again simulate two independent components and take a linear combination to achieve a certain desired correlation. The difference in this nonsta-

tionary setting is that for each simulation we also randomly select the parameter, $\tau \in \{100, 412\}$ and only take a linear combination for the time points, $t \in \{\tau, T-1\}$. The correlation will therefore be a step function, initially equal to zero and then rising to a value, r, at time point $t = \tau$. For each simulation we apply the MvLSW test to all time points. The proportion of positive results for the time points $t \in \{0, \tau - 1\}$ is the false discovery rate for that simulation, similarly the proportion of positive results for the time points $t \in \{\tau, T-1\}$ is the power for that simulation. We repeat this for 1000 simulations and average the power and false discovery rates.

The results from the nonstationary simulation study are shown in Table 5.3. We see from the results that the true discovery rate is similar to the stationary case indicating the test is still powerful despite the nonstationary structure being considered. The false discovery rate is however slightly higher, most notably for models **N1** and **N2** where it can be as high as 0.136.

5.5 EEG Example

Returning to the EEG example from Section 5.2 we perform our coherence test on the three channels. Again we controlled the expected false discovery rate to be below 0.05. The results are shown in Figure 5.5. Significant coherence was detected between all pairs of channels and in all cases this is driven by level $j = 3$. The pair C4 and FC4 are significantly coherent for only a short span of time points between the times $t/T = 0.2$ and $t/T = 0.4$. The pair FC4 and FC6 on the other hand show significant coherence for almost the entire time span. Finally the pair C4 and FC6 show significant coherence

Model	r	MvLSW	
		power	false
N1	0.9	0.976	0.135
	0.8	0.945	0.103
	0.7	0.898	0.075
N2	0.9	0.984	0.136
	0.8	0.967	0.108
	0.7	0.919	0.085
N3	0.9	0.928	0.092
	0.8	0.868	0.068
	0.7	0.754	0.048

Table 5.3: Nonstationary Simulation Study Results

for most of the first half of the time span. It is interesting to note that both pairs C4 and FC4 and C4 and FC6 stop showing significant coherence close to the time point where the stimulus is given to the participant. These results are not unexpected given our earlier analysis of the confidence intervals.

5.6 Conclusion and Discussion

In conclusion we have shown that for a multivariate series represented under the MvLSW model it is possible to construct an approximate confidence interval for the estimated LWS matrix. Calculating this confidence interval can give a clear indica-

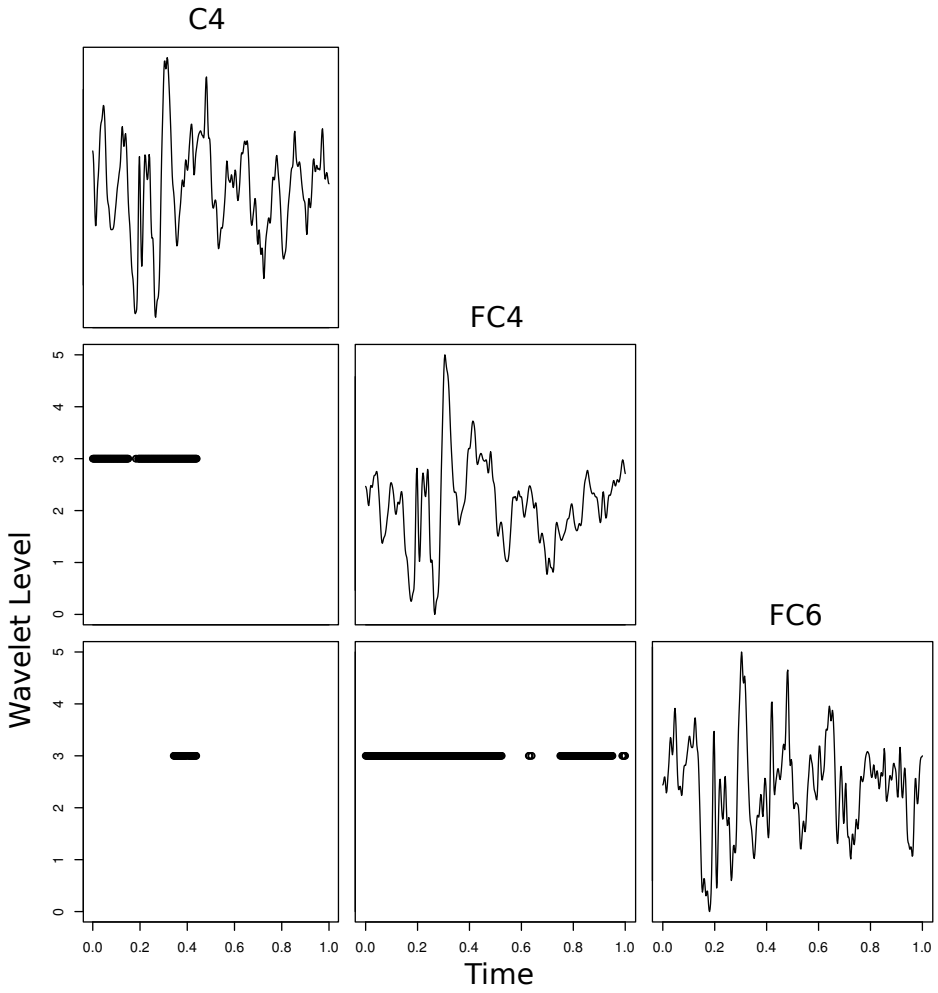


Figure 5.4: Results of the test of coherence for EEG data. Plots on the diagonal show the recordings themselves. The off diagonal plots show the scale and location points which are found to be significantly coherent.

tion of which elements of the true LWS are likely to be nonzero. We demonstrated the accuracy of our confidence interval by comparing it to one calculated using bootstrapping. A confidence interval was also calculated for the LWS estimate for an EEG recording. Leading on from the calculation of confidence intervals we defined a hypothesis test for coherence. We compared this test to a Fourier bases test in a simulation study for stationary time series models. The MvLSW test was also applied to

nonstationary time series models. Finally we applied our MvLSW test of coherence to EEG data to show how it can be used in practice to identify significant dependencies between channels.

Chapter 6

Conclusions and Discussion of Future Work

Wavelet methods are now well established in the field of time series analysis. The locally stationary wavelet model (LSW) provides a wavelet based representation of a time series without assuming second order stationarity. In Chapter 3 we introduced the multivariate locally stationary wavelet model (MvLSW) which generalises the univariate LSW model to cover multivariate time series. We also demonstrated how the MvLSW model is able to effectively model the dependencies between components of a multivariate series via wavelet coherence and wavelet partial coherence. Just as the LSW model does not assume constant auto-covariance the MvLSW model does not assume second order stationarity both in terms of the auto-covariance structure of the individual components *and* the cross-covariance structure between components.

In Chapter 4 we introduced an application of the MvLSW model to the problem of time series classification. The specific classification problem which we look at is the

dynamic classification problem. What makes this different to the usual time series classification case is that the class membership of a series is permitted to change over time within a pre-determined set of potential classes. Our proposed approach is to first estimate the coherence of a series with unknown class membership. At each location we compare the estimated coherence to the coherence structure of each class in order to calculate the probability of the series being in a particular class at that location. Our approach is shown to work well on both simulated and accelerometer data.

Finally Chapter 5 introduced a second application of the MvLSW model namely a test of coherence. Testing the coherence between components of a multivariate time series is important as it enables us to identify which are significantly coherent at a given time. It also identifies which wavelet levels are driving any coherence.

We conclude this thesis by considering some possible future developments of this research. We first consider the MvLSW model introduced in Chapter 3, an interesting extension to this model would be to permit the use of the wavelet packet transform rather than the non-decimated wavelet transform. The wavelet packet transform is discussed in Section 2.4.3 and allows for greater flexibility in the model selection. Cardinali and Nason (2008) consider the use of wavelet packets in the univariate LSW setting but this has so far not been considered in a multivariate time series setting.

If we consider the dynamic classification method described in Chapter 4 one possibile future direction is to consider a scenario where the set of training data is incomplete. For example the training data may contain examples of N known classes, however there may be additional, unobserved classes. Such a scenario would require a

method which can distinguish between time points belonging to one of the N observed classes and those belonging to an unobserved class. Alternatively we may consider a scenario where there is no training data available. Solving this problem would require an unsupervised classifiction approach to estimate both the number and structure of the classes.

Finally we consider possible extensions of the test of coherence described in Chapter 5. In Section 5.3 we noted that the possibility of dependencies between test statistics is not accounted for in the FDR procedure. Benjamini and Yekutieli (2001) show that, under some mild assumptions about the dependencies, the FDR procedure will still controll the false discovery rate but may be more conservative. A useful extension to our work would be to establish if the assumptions made by Benjamini and Yekutieli (2001) are satisfied in our setting and if we can guard against the test being overly conservative.

Another interesting extension would be to construction a test of partial coherence. Partial coherence is able to distinguish between components which are directly dependent and those which are dependent only through dependencies with other components. A test of partial coherence would therefore enable us to formally distinguish between these two scenarios. Early investigation into this subject indicates that it is difficult to construct a suitable test statistic with a known distribution. The ability to accurately map out the dependence structure also raises the possibility of dimension reduction. Methods discussed in Jolliffe (1972) and McCabe (1984) seek to select a subset of channels whilst retaining the majority of the information. Knowledge of the (time varying) dependencies could then be incorporated into a similar scheme.

Appendix A

Proofs for Chapter 3

A.1 Proof of Proposition 3.1

Suppose, by way of contradiction, that there exist two representations for the same process, $\mathbf{V}_j^{(1)}(u)$ and $\mathbf{V}_j^{(2)}(u)$. At each time point, u, there exists $\mathbf{S}_j^{(1)}(u)$ and $\mathbf{S}_j^{(2)}(u)$ such that,

$$\mathbf{c}(u,\tau) = \sum_{j=1}^{\infty} \mathbf{S}_j^{(1)}(u)\Psi_j(\tau) = \sum_{j=1}^{\infty} \mathbf{S}_j^{(2)}(u)\Psi_j(\tau). \tag{A.1}$$

Let $\mathbf{\Delta}_j(u)$ be a matrix representing the element-wise difference between the two representations, From equation (A.1) it is clear that,

$$\sum_{j=1}^{\infty} \mathbf{\Delta}_j(u)\Psi_j(\tau) = \mathbf{0}, \quad \forall u \in (0,1) \text{ and } \tau \in \mathbb{Z}. \tag{A.2}$$

To establish the uniqueness of the MvLWS representation we must show that (A.2) implies that, $\mathbf{\Delta}_j(u) = \mathbf{0}\ \forall j > 0,\ u \in (0,1)$. Using arguments similar to those set out by Nason et al. (2000) we use Parseval's relation and the definition of the inner product matrix to obtain, $A_{jl} = \sum_\tau \Psi_j(\tau)\Psi_l(\tau) = \frac{1}{2\pi}\int d\omega \hat{\Psi}_j(\omega)\hat{\Psi}_l(\omega)$, where $\hat{\Psi}_j(\omega) =$

$\left|\hat{\psi}_j(\omega)\right|^2 = 2^j \left|m_1(2^{(j+1)}\omega)\right|^2 \prod_{l=0}^{(j-2)} \left|m_0(2^l\omega)\right|^2$, and $m_0(\omega) = 2^{-1/2}\sum_k h_k \exp(-i\omega k)$, with $\sum_k h_k^2 = 1$, $\frac{1}{\sqrt{2}}\sum_k h_k = 1$ and $|m_1(\omega)|^2 = 1 - |m_0(\omega)|^2$. From equation (A.2) we can say that for a general element:

$$\sum_l \sum_j \Delta_j^{(p,q)}(u)\Delta_l^{(p,q)}(u) \sum_\tau \Psi_j(\tau)\Psi_l(\tau) = 0$$

Hence it is easily shown that,

$$\int d\omega \left(\sum_j \Delta_j^{(p,q)}(u)\hat{\Psi}_j(\omega)\right)^2 = 0. \tag{A.3}$$

Since we have already made the assuption that, $\sum_j S_j^{(p,q)} < \infty \ \forall p, q$, we infer that $\sum_j \Delta_j^{(p,q)}(u)\hat{\Psi}_j(\omega)$ is continuous in $\omega \in [-\pi, \pi]$, because every $\hat{\Psi}_j(\omega)$ is and $\sum_j \left|\Delta_j^{(p,q)}(u)\right| < \infty$. Hence (A.3) implies that, $\sum_{j=1}^{\infty} \mathbf{\Delta}_j(u)\hat{\Psi}_j(\omega) = \mathbf{0}$. The remainder of the proof then follows similarly to Nason et al. (2000).

■

A.2 Proof of Proposition 3.2

Recall the definition of the wavelet representation of a multivariate series in equation (3.2).

$$\begin{aligned}
\operatorname{cov}\left(X_{uT}^{(p)}, X_{uT+\tau}^{(q)}\right) &= E\left[X_{uT}^{(p)} X_{uT+\tau}^{(q)}\right], \\
&= \mathrm{E}\left[\sum_{j=1}^{\infty}\sum_{k}\sum_{r=1}^{p} V_j^{(p,r)}(k/T)\psi_{j,k}(uT) z_{j,k}^{(r)}\right. \\
&\left. \times \sum_{j'=1}^{\infty}\sum_{k'}\sum_{r'=1}^{q} V_{j'}^{(q,r')}(k'/T)\psi_{j',k'}(uT+\tau) z_{j',k'}^{(r')}\right], \\
&= \sum_{j=1}^{\infty}\sum_{k}\sum_{r=1}^{\min p,q} V_j^{(p,r)}(k/T)V_j^{(q,r)}(k/T)\psi_{jk}(uT)\psi_{jk}(uT+\tau).
\end{aligned}$$

Recalling the definition of the LWS matrix we can say that,

$S_j^{(p,q)}(u) = \sum_{r=1}^{\min p,q} V_j^{(p,r)}(u) V_j^{(q,r)}(u)$. We also make the substitition $m = k - uT$ to obtain,

$$\operatorname{cov}\left(X_{uT}^{(p)}, X_{uT+\tau}^{(q)}\right) = \sum_j \sum_m S_j\left(\frac{uT+m}{T}\right)\psi_{jm}(0)\psi_{jm}(\tau).$$

Analogous to the approach considered by Nason et al. (2000) in the univariate setting, using the assumed Lipschitz continuous property of $V_j^{(p,q)}(z)$ and therefore $S_j^{(p,q)}(z)$we can consider the difference between this covariance and the function $c^{(p,q)}(u,\tau)$,

$$\begin{aligned}\left|\operatorname{cov}\left(X_{uT}^{(p)}, X_{uT+\tau}^{(q)}\right) - c^{(p,q)}(u,\tau)\right| &= \left|\sum_j \sum_m S_j\left(\frac{uT+m}{T}\right)\psi_{jm}(0)\psi_{jm}(\tau) - c^{(p,q)}(u,\tau)\right| \\ &\leq T^{-1}\sum_m |m|\, L_j\, |\psi_{jm}(0)\psi_{jm}(\tau)| = \mathcal{O}(T^{-1}).\end{aligned}$$

■

A.3 Proof of Proposition 3.3

To establish this result we firstly demonstrate that $\mathbf{S}_j^*(u)$ is positive definite. Since $\mathbf{S}_j(u)$ is positive definite, by Choleski, there exists a lower triangular matrix $\mathbf{V}_j(u)$ so that $\mathbf{S}_j(u) = \mathbf{V}_j(u)\mathbf{V}_j'(u)$. Hence $\mathbf{S}_j^*(u) = \mathbf{M}\mathbf{V}_j(u)\mathbf{V}_j'(u)\mathbf{M}' = (\mathbf{M}\mathbf{V}_j(u))(\mathbf{M}\mathbf{V}_j(u))'$. Hence $\mathbf{S}_j^*(u)$ is positive definite. Second, since $\mathbf{S}_j^*(u)$ is positive definite, there exists a lower triangular matrix $\mathbf{V}_j^*(u)$ such that $\mathbf{S}_j^*(u) = \mathbf{V}_j^*(u)\mathbf{V}_j'^*(u)$. Thus $\mathbf{X}_t^*$ admits a MvLSW representation with transfer function $\mathbf{V}_j^*(u)$.

■

A.4 Proof of Proposition 3.4

Expectation: Recall that $d_{j,k}^{(p)} = \sum_t X_t^{(p)} \psi_{j,k}(t)$ and

$X_t^{(p)} = \sum_l \sum_m \sum_r V_l^{(p,r)}(m/T)\psi_{l,m}(t) z_{l,m}^{(r)}$. Hence

$$\begin{aligned} E\left[I_{j,k}^{(p,q)}\right] =& E\left[\left\{\sum_t X_t^{(p)}\psi_{j,k}(t)\right\}\left\{\sum_{t'} X_{t'}^{(q)}\psi_{j,k}(t')\right\}\right], \\ =& \sum_{l=1}^{J}\sum_m \sum_{r=1}^{\min\{p,q\}} V_l^{(p,r)}(m/T)\; V_l^{(q,r)}(m/T) \times \left\{\sum_t \psi_{l,m}(t)\psi_{j,k}(t)\right\}^2. \end{aligned} \tag{A.4}$$

Substituting $m = n + k$ into (A.4) we obtain,

$$E\left[I_{j,k}^{(p,q)}\right] = \sum_{l=1}^{J}\sum_n \left\{S_l^{(p,q)}\left(\frac{n+k}{T}\right)\right\}\left\{\sum_t \psi_{l,n+k-t}\psi_{j,k-t}\right\}^2.$$

Analogous to the univariate setting of Nason et al. (2000), since $S_j^{(p,q)}(z)$, is Lipschitz continuous with finite Lipschitz constant L_j, for some fixed n, $\left|S_j^{(p,q)}((k+n)/T) - S_j^{(p,q)}(k/T)\right| \leq |n|\, L_j/T$, and therefore $S_j^{(p,q)}((n+k)/T) = S_j^{(p,q)}(k/T) + \mathcal{O}(T^{-1})$. Consequently

$$\begin{aligned} E\left[I_{j,k}^{(p,q)}\right] =& \sum_{l=1}^{J} S_l^{(p,q)}\left(\frac{k}{T}\right)\sum_t\sum_v \psi_{j,-t}\psi_{j,-v-t} \\ & \times \sum_n \psi_{l,n-t}\psi_{l,n-v-t} + \mathcal{O}(T^{-1}). \end{aligned} \tag{A.5}$$

Recalling the definition of the autocorrelation wavelets we find that,

$$\begin{aligned} E\left[I_{j,k}^{(p,q)}\right] &= \sum_{l=1}^{J} S_l^{(p,q)}\left(\frac{k}{T}\right)\sum_v \Psi_l(v)\Psi_j(v) + \mathcal{O}(T^{-1}), \\ &= \sum_{l=1}^{J} A_{jl} S_l^{(p,q)}\left(\frac{k}{T}\right) + \mathcal{O}(T^{-1}). \end{aligned}$$

■

Variance: To establish the variance of the raw periodogram, we begin by considering $E\left[(I_{j,k}^{(p,q)})^2\right] = E\left[\left(d_{j,k}^{(p)}\right)^2\left(d_{j,k}^{(p)}\right)^2\right]$.

$$\begin{aligned}E\left[(I_{j,k}^{(p,q)})^2\right] = &\left(\sum_{l=1}^{J}\sum_{m}\sum_{r=1}^{p} V_l^{(p,r)}(m/T)\sum_t \psi_{l,m}(t)\psi_{j,k}(t)\right.\\ &\left.\times\sum_{l'=1}^{J}\sum_{m'}\sum_{r'=1}^{q} V_{l'}^{(q,r')}(m'/T)\sum_{t'}\psi_{l',m'}(t')\psi_{j,k}(t')\right)^2\\ &\times E\left[z_{l_1,m_1}^{(r_1)} z_{l_2,m_2}^{(r_2)} z_{l_3,m_3}^{(r_3)} z_{l_4,m_4}^{(r_4)}\right].\end{aligned}$$

Using a result due to Isserlis (1918) the above expression can be re-written as the sum of three different elements $E\left[(I_{j,k}^{(p,q)})^2\right] = I_1 + I_2 + I_3$ where, for example,

$$I_1 = \prod_{i=1}^{4}\sum_{t_i,l_i,m_i,r_i} V_{l_i}^{(p_i,r_i)}(m_i/T)\psi_{l_i m_i}(t_i)\psi_{jk}(t_i) E\left[z_{l_1,m_1}^{(r_1)} z_{l_2,m_2}^{(r_2)}\right] E\left[z_{l_3,m_3}^{(r_3)} r_{l_4,m_4}^{(r_4)}\right].$$

Since $E\left[z_{l_1,m_1}^{(r_1)} z_{l_2,m_2}^{(r_2)}\right] = \delta_{l_1 l_2}\delta_{m_1 m_2}\delta_{r_1 r_2}$ this simplifies to:

$$\begin{aligned}I_1 = &\sum_{l_1,m_1,r_1}\left(V_{l_1}^{(p,r_1)}(m_1/T)\right)^2 \times \sum_{t_1=0}^{T-1}\psi_{l_1,m_1}(t_1)\psi_{j,k}(t_1)\\ &\times\sum_{t_2=0}^{T-1}\psi_{l_1,m_1}(t_2)\psi_{j,k}(t_2)\sum_{l_3,m_3,r_3}\left(V_{l_3}^{(q,r_3)}(m_3/T)\right)^2\\ &\times\sum_{t_3=0}^{T-1}\psi_{l_3,m_3}(t_3)\psi_{j,k}(t_3)\sum_{t_4=0}^{T-1}\psi_{l_3,m_3}(t_4)\psi_{j,k}(t_4);\\ &= E\left[I_{j,k}^{(p,p)}\right] E\left[I_{j,k}^{(q,q)}\right].\end{aligned}$$

Similarly for I_2 we find that $I_2 = E\left[I_{j,k}^{(p,q)}\right]^2$ and $I_3 = E\left[I_{j,k}^{(p,q)}\right]^2$. Hence,

$$E\left[(I_{j,k}^{(p,q)})^2\right] = E\left[I_{j,k}^{(p,p)}\right] E\left[I_{j,k}^{(q,q)}\right] + 2E\left[I_{j,k}^{(p,q)}\right]^2,$$

and

$$\begin{aligned}\mathrm{Var}\left\{I_{j,k}^{(p,q)}\right\} &= \left(\sum_{l=1}^{J} A_{jl} S_l^{(p,p)}\left(\frac{k}{T}\right) + \mathcal{O}(T^{-1})\right) \\ &\times \left(\sum_{l=1}^{J} A_{jl} S_l^{(q,q)}\left(\frac{k}{T}\right) + \mathcal{O}(T^{-1})\right) \\ &+ \left(\sum_{l=1}^{J} A_{jl} S_l^{(p,q)}\left(\frac{k}{T}\right) + \mathcal{O}(T^{-1})\right)^2.\end{aligned}$$

From Nason et al. (2000) it is known that $\sum_\tau |\Psi_j(\tau)| = \mathcal{O}(2^j)$, and hence $A_{jl} = \sum_\tau \Psi_j(\tau)\Psi_l(\tau) \leq \left(\sum_\tau |\Psi_j(\tau)|\right)^2 = \mathcal{O}(2^{2j})$. Hence it is easily verified that,

$$\begin{aligned}\mathrm{Var}\left\{I_{j,kT}^{(p,q)}\right\} =& \sum_{l=1}^{J} A_{jl} S_l^{(p,p)}\left(\frac{k}{T}\right) \sum_{l=1}^{J} A_{jl} S_l^{(q,q)}\left(\frac{k}{T}\right) \\ &+ \left(\sum_{l=1}^{J} A_{jl} S_l^{(p,q)}\left(\frac{k}{T}\right)\right)^2 + \mathcal{O}(2^{2j}/T).\end{aligned}$$

■

A.5 Proof of Proposition 3.5

Recall that the form of the smoothed periodogram is, $\tilde{\mathbf{I}}_{j,k} = (2M+1)^{-1} \sum_{m=-M}^{M} \mathbf{I}_{j,k+m}$.

Expectation:

$$E\left[\tilde{I}_{j,k}^{(p,q)}\right] = \frac{1}{2M+1} \sum_{m=-M}^{M} E\left[I_{j,k+m}^{(p,q)}\right].$$

Where $2M+1$ is the size of the smoothing window. Using the expected value of the periodogram previously calculated this becomes,

$$E\left[\tilde{I}_{j,k}^{(p,q)}\right] = \frac{1}{2M+1} \sum_{m=-M}^{M} \sum_{l=1}^{J} \left\{A_{jl} S_l^{(p,q)}\left(\frac{k+m}{T}\right) + \mathcal{O}(T^{-1})\right\}.$$

Due to the Lipschitz continuity assumed for the spectral components it follows that:

$$E\left[\tilde{I}_{j,k}^{(p,q)}\right] = \sum_{l=1}^{J} A_{jl} S_l^{(p,q)}\left(\frac{k}{T}\right) + \mathcal{O}(MT^{-1}).$$

As $T \to \infty$, $M \to \infty$ but $\frac{M}{T} \to 0$, the smoothed raw wavelet periodogram (auto and cross) is asymptotically biased in the usual way. As such it can be corrected by use of the inverse inner product matrix, A^{-1} to achieve an asymptotically unbiased estimate.

■

Variance: We begin by considering: $E\left[\left(\tilde{I}_{j,k}^{(p,q)}\right)^2\right]$.

$$E\left[\left(\tilde{I}_{j,k}^{(p,q)}\right)^2\right] = \frac{1}{(2M+1)^2} \sum_{m=-M}^{M} \sum_{m'=-M}^{M} E\left[I_{j,k+m}^{(p,q)} I_{j,k+m'}^{(p,q)}\right],$$

by substituting $\tau = m' - m$. Using arguments similar to those employed in the proof of the Expectation, it follows that:

$$\frac{1}{(2M+1)^2} \sum_{m=-M}^{M} \sum_{\tau=M-m}^{M+m} E\left[I_{j,k+m}^{(p,q)} I_{j,k+m+\tau}^{(p,q)}\right] = \frac{1}{(2M+1)^2} \sum_{m,\tau} E\left[d_{j,k+m}^{(p)} d_{j,k+m}^{(q)} d_{j,k+m+\tau}^{(p)} d_{j,k+m+\tau}^{(q)}\right],$$

Using Isserlis' Theorem Isserlis (1918), it can be shown that

$$\begin{aligned}
\mathrm{Var}\left\{\tilde{I}_{j,k}^{(p,q)}\right\} =& \frac{1}{(2M+1)^2}\left\{\sum_{m,\tau} E\left[d_{j,k+m}^{(p)} d_{j,k+m+\tau}^{(p)}\right] E\left[d_{j,k+m}^{(q)} d_{j,k+m+\tau}^{(q)}\right]\right. \\
&\left. + \sum_{m,\tau} E\left[d_{j,k+m}^{(p)} d_{j,k+m+\tau}^{(q)}\right] E\left[d_{j,k+m}^{(q)} d_{j,k+m+\tau}^{(p)}\right]\right\}, \\
=& \frac{1}{(2M+1)^2} \sum_{m=-M}^{M}\left\{\sum_{\tau}\sum_{l=1}^{J} S_l^{(p,p)}(k/T) A_{l,j}^{\tau}\right. \\
&\times \sum_{l'=1}^{J} S_{l'}^{(q,q)}(k/T) A_{l',j}^{\tau} + \sum_{\tau}\left(\sum_{l=1}^{J} S_l^{(p,q)}(k/T) A_{l,j}^{\tau}\right)^2 \\
&\left. + \sum_{\tau}(|m|+1)\mathcal{O}(T^{-1}) + \sum_{\tau}(|m|+1)^2\mathcal{O}(T^{-2})\right\}.
\end{aligned}$$

where $A^{\tau}_{l,j} = \sum_t \Psi_{l,j}(t)\Psi_{l,j}(t+\tau)$. Note that this is a form of inner product matrix but with a given lag, τ. Examining the term,

$$\begin{aligned}
&\sum_{\tau}\sum_{l=1}^{J} S_l^{(p,p)}\left(k/T\right) A^{\tau}_{l,j} \sum_{l'=1}^{J} S_{l'}^{(q,q)}\left(k/T\right) A^{\tau}_{l',j} \\
&\qquad \leq \left(\sum_{\tau}\left|\sum_{l=1}^{J} S_l^{(p,p)}\left(k/T\right) A^{\tau}_{l,j}\right|\right)\left(\sum_{\tau}\left|\sum_{l'=1}^{J} S_{l'}^{(q,q)}\left(k/T\right) A^{\tau}_{l',j}\right|\right), \\
&\qquad = \left(\sum_{n}\left|c^{(p,p)}(k,n)\right|\sum_{\tau}\left|\Psi_{l,j}(n+\tau)\right|\right) \\
&\qquad \times \left(\sum_{n}\left|c^{(q,q)}(k,n)\right|\sum_{\tau}\left|\Psi_{l,j}(n+\tau)\right|\right) = \mathcal{O}(2^{2j}).
\end{aligned}$$

Similarly it can be shown that the second term is also equal to $\mathcal{O}(2^{2j})$ hence,

$$\mathrm{Var}\left\{\tilde{I}^{(p,q)}_{j,k}\right\} = \mathcal{O}(2^{2j}/M) + \mathcal{O}(2^{2j}/T). \tag{A.6}$$

Thus, the smoothed wavelet auto and cross periodogram is asympotically mean-squared consistent as $T \to \infty$, $M \to \infty$, $\frac{M}{T} \to 0$.

■

Appendix B

Proofs for Chapter 4

B.1 Proof of Proposition 4.1

We begin by reminding the reader of a result established by Park et al. (2014) which is relevant to this proof, namely that the variance of the LWS estimate, $\widehat{S}_{jk}^{(p,q)}$, can be expressed as,

$$\text{Var}\left\{\widehat{S}_{jk}^{(p,q)}\right\} = \mathcal{O}(M_T^{-1}) + \mathcal{O}(T^{-1}).$$

Here M_T is the smoothing bandwidth used to calculate $\widehat{S}_{jk}^{(p,q)}$. For this estimate to be both asymptotically unbiased and consistent Park et al. (2014) make the assumptions that $M_T \to \infty$ and $M_T/T \to 0$ in the limit as $T \to \infty$. Given this we can express M_T in the form $M_T = \mathcal{O}(T^{\alpha})$ for some $\alpha \in (0, 1)$. The variance of $\widehat{S}_{jk}^{(p,q)}$ can then be expressed as a single order term, $\text{Var}\left\{\widehat{S}_{jk}^{(p,q)}\right\} = \mathcal{O}(T^{-\alpha})$.

We now consider the asymptotics of our classification procedure. Let $\widehat{\boldsymbol{\mu}}_k$ be a vector of length N which contains the elements of $\widehat{\boldsymbol{\zeta}}_{j,k;X}$ which will be used to distinguish

the different classes. For simplicity we will consider the two class problem however the results are easily generalised to the more general case. We define the divergence criterion to be,

$$\Delta(\widehat{\boldsymbol{\mu}}_k) = \frac{1}{2}\left\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2) - (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_1) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right\}. \tag{B.1}$$

This divergence criterion is simply the difference between log-likelihoods under the two classes. We also define the classification decision rule,

$$D(\widehat{\boldsymbol{\mu}}_k) = \begin{cases} 1 & (\text{estimate } C(k) = 1) \text{ if } \quad \Delta(\widehat{\boldsymbol{\mu}}_k) > 0 \\ 2 & (\text{estimate } C(k) = 2) \text{ if } \quad \Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0 \end{cases}.$$

Suppose that the true class membership, $C(k)$, is equal to 1. Here we want to show that the probability of misclassification goes to 0 as $T \to \infty$. That is we want to show, $\Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1) \to 0$, or equivalently, $\Pr(\Delta(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k) = 1) \to 0$.

What we will actually show is that for the scaled divergence, $\delta_T(\widehat{\boldsymbol{\mu}}_k) = \Delta(\widehat{\boldsymbol{\mu}}_k)/T^\alpha$ for some $\alpha \in (0, 1)$, that $\Pr(\delta_T(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k) = 1) \to 0$ as $T \to \infty$, and consequently that $\Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1) \to 0$ in the same limit. This results immediately follows if we can establish that as $T \to \infty$ then $\delta_T(\widehat{\boldsymbol{\mu}}_k) \xrightarrow{P} K \geq 0$, which is satisfied by the following two conditions in the limit a $T \to \infty$: **A1:** $E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] \to K$ where $K > 0$ and **A2:** $\text{Var}\left\{\delta_T(\widehat{\boldsymbol{\mu}}_k)\right\} \to 0$,

Expectation of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$

We first consider the expectation of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$,

$$E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = -\frac{1}{2T^\alpha}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\right] + \frac{1}{2T^\alpha}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right] + \frac{1}{2T^\alpha}E\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right].$$

We note that the first term follows a chi-squared distribution with N degrees of freedom, the expectation of which is equal to N. We now focus on the second term,

$$\begin{aligned}
E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right] =& E\left[\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\}'\boldsymbol{\Sigma}_2^{-1}\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\}\right], \\
=& E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\right] + 2E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right] \\
&+ E\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]. \\
=& E\left[\mathrm{tr}\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\}\right] + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2), \\
=& E\left[\mathrm{tr}\{\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\}\right] + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2), \\
=& \mathrm{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2). \qquad \text{(B.2)}
\end{aligned}$$

Therefore,

$$E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2T^\alpha}\left\{-N + \mathrm{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right\}.$$

Using the results of Proposition 5 from Park et al. (2014), the variance covariance matrices can be expressed as,

$$\begin{aligned}
\boldsymbol{\Sigma}_c =& \frac{\boldsymbol{A}_c}{T^\alpha}, \\
\boldsymbol{\Sigma}_c^{-1} =& \boldsymbol{B}_c T^\alpha.
\end{aligned}$$

Here $\boldsymbol{A}_c$ and $\boldsymbol{B}_c$ are constant symmetric positive definite matrices. The expectation can then be written as,

$$E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2T^\alpha}\left\{-N + \mathrm{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\} + T^\alpha(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{B}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \log\frac{|\boldsymbol{A}_2|}{|\boldsymbol{A}_1|}\right\}.$$

Since $\boldsymbol{B}_2$ is positive definite then we can say $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{B}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = C$ for some constant $C > 0$. Also, since $\boldsymbol{A}_1$ and $\boldsymbol{B}_2$ are symmetric and positive definite, Meenakshi and Rajian (1999) showed that $\boldsymbol{A}_1\boldsymbol{B}_2$ will be positive definite and so we can say that $\mathrm{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\} = D$ for some $D > 0$. Finally the term $|\boldsymbol{A}_2|/|\boldsymbol{A}_1|$ must be positive as both $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are positive definite. Without loss of generality we assume that this term is equal to $G \in (0,\infty)$. We can therefore express the expectation as,

$$\begin{aligned} E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] =& \frac{1}{2T^\alpha}\left\{-N + CT^\alpha + D + \log G\right\}, \\ =& \frac{C}{2} + \frac{D - N + \log G}{2T^\alpha}. \end{aligned}$$

Clearly as $T \to \infty$ then $E\left[\delta_T(\widehat{\boldsymbol{\mu}}_k)\right] \to C/2$ where $C/2$ is a positive constant therefore condition **A1** is satisfied.

Variance of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$

We now consider the variance of $\delta_T(\widehat{\boldsymbol{\mu}}_k)$,

$$\begin{aligned} \mathrm{Var}\left\{\delta_T(\widehat{\boldsymbol{\mu}}_k)\right\} =& \frac{1}{4T^{2\alpha}}\mathrm{Var}\left\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\right\} + \frac{1}{4T^{2\alpha}}\mathrm{Var}\left\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right\} \\ &+ \frac{1}{2T^{2\alpha}}\mathrm{cov}\left((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1), (\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right). \end{aligned} \tag{B.3}$$

The first term is simply the variance of a chi-squared random variable with N degrees of freedom so is equal to $2N$. We therefore focus on the second and third terms.

Looking at the second term,

$$\begin{aligned}\mathrm{Var}\left\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right\} =&E\left[\left\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right\}^2\right]\\ &-\left\{E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right]\right\}^2.\end{aligned}$$

The second term in the above equation, $\left\{E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right]\right\}^2$, is simply the square of the term found in equation (B.2). We therefore focus on the first term, $E\left[\left\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right\}^2\right]$. For simplicity we make the substitution $\boldsymbol{\mu}^*=(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)$,

$$\begin{aligned}E\left[\left\{\boldsymbol{\mu}^{*\prime}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}^*\right\}^2\right] =&E\left[\boldsymbol{\mu}^{*\prime}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*\prime}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}^*\right],\\ =&E\left[\sum_i\sum_j\boldsymbol{\mu}_i^*\left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\boldsymbol{\mu}_j^*\sum_{i'}\sum_{j'}\boldsymbol{\mu}_{i'}^*\left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'}\boldsymbol{\mu}_{j'}^*\right],\\ =&\sum_i\sum_j\sum_{i'}\sum_{j'}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'}E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_j^*\boldsymbol{\mu}_{i'}^*\boldsymbol{\mu}_{j'}^*\right],\\ =&\sum_i\sum_j\sum_{i'}\sum_{j'}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij}\left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'}\left\{E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_j^*\right]E\left[\boldsymbol{\mu}_{i'}^*\boldsymbol{\mu}_{j'}^*\right]\right.\\ &\left.+E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_{i'}^*\right]E\left[\boldsymbol{\mu}_j^*\boldsymbol{\mu}_{j'}^*\right]+E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_{j'}^*\right]E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_{j'}^*\right]\right\}.\end{aligned}$$

In the final step above we have used Isserlis' theorem, Isserlis (1918), to split the expression into three terms. We label these as D_1, D_2 and D_3. We also note that $E\left[\boldsymbol{\mu}_i^*\boldsymbol{\mu}_j^*\right]=(\boldsymbol{\Sigma}_1)_{ij}+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)_j$. Looking at these terms individually we

have,

$$
\begin{aligned}
D_1 =& \sum_{iji'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} E\left[\boldsymbol{\mu}_i^* \boldsymbol{\mu}_j^*\right] E\left[\boldsymbol{\mu}_{i'}^* \boldsymbol{\mu}_{j'}^*\right], \\
=& \sum_{iji'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} \Big\{ (\boldsymbol{\Sigma}_1)_{ij} (\boldsymbol{\Sigma}_1)_{i'j'} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} (\boldsymbol{\Sigma}_1)_{ij} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} \Big\}, \\
=& \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\Sigma}_1)_{ij} \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\Sigma}_1)_{i'j'} \\
& + \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\Sigma}_1)_{ij} \\
& + \sum_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j \sum_{i'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'}, \\
=& \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\}^2 + 2\ \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2, \\
=& E\left[(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right]^2.
\end{aligned}
$$

Following a similar procedure for D_2,

$$
\begin{aligned}
D_2 =& \sum_{iji'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} E\left[\boldsymbol{\mu}_i^* \boldsymbol{\mu}_{i'}^*\right] E\left[\boldsymbol{\mu}_j^* \boldsymbol{\mu}_{j'}^*\right], \\
=& \sum_{iji'j'} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{ij} \left(\boldsymbol{\Sigma}_2^{-1}\right)_{i'j'} \Big\{ (\boldsymbol{\Sigma}_1)_{ii'} (\boldsymbol{\Sigma}_1)_{jj'} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\Sigma}_1)_{jj'} \\
& + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} (\boldsymbol{\Sigma}_1)_{ii'} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_j (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{i'} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{j'} \Big\}, \\
=& \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}^2.
\end{aligned}
$$

Similarly $D_3 = D_2$. Putting together we obtain,

$$
\begin{aligned}
\text{Var}\left\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_2)\right\} =& 2\text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} + 4(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
& + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).
\end{aligned}
$$

We now consider the covariance term in equation (B.3),

$$\begin{aligned}\operatorname{cov}\left((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1),(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right)=&\\ E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right]&\\ -E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)\right]E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right].&\end{aligned}$$

Looking at the first term,

$$\begin{aligned}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right]=&\\ \sum_{iji'j'}(\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i'j'}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_i(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_j(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{i'}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{j'}\right].&\end{aligned}$$

We again split this up into three terms, C_1, C_2 and C_3,

$$\begin{aligned}C_1=&\sum_{iji'j'}(\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i'j'}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_i(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_j\right]E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{i'}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{j'}\right],\\ =&\sum_{iji'j'}(\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i'j'}\left\{(\boldsymbol{\Sigma}_1)_{ij}(\boldsymbol{\Sigma}_1)_{i'j'}+(\boldsymbol{\Sigma}_1)_{ij}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)_{i'}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)_{j'}\right\},\\ =&N\operatorname{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}+N(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).\end{aligned}$$

Similarly for C_2,

$$\begin{aligned}C_2=&\sum_{iji'j'}(\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i'j'}E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_i(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{i'}\right]E\left[(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)_j(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)_{j'}\right],\\ =&\sum_{iji'j'}(\boldsymbol{\Sigma}_1^{-1})_{ij}(\boldsymbol{\Sigma}_2^{-1})_{i'j'}(\boldsymbol{\Sigma}_1)_{ii'}(\boldsymbol{\Sigma}_1)_{jj'},\\ =&\operatorname{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}=\operatorname{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}.\end{aligned}$$

It can also be shown that $C_2=C_3$ therefore,

$$\operatorname{cov}\left((\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1),(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)\right)=2\operatorname{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}.$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left\{\delta_T(\widehat{\boldsymbol{\mu}}_k)\right\} =& \frac{1}{2T^{2\alpha}}\left[N + \mathrm{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} + 2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right.\\ &\left.+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \mathrm{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}\right],\\ =& \frac{1}{2T^{2\alpha}}\left[N + \mathrm{tr}\left\{\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2\boldsymbol{A}_1\right\} + 2T^{\alpha}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right.\\ &\left.+\ T^{\alpha}(\boldsymbol{\mu_1}-\boldsymbol{\mu}_2)'\boldsymbol{B}_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + 4\mathrm{tr}\{\boldsymbol{B}_1\boldsymbol{A}_2\}\right].\end{aligned}$$

Using similar arguments as for the expectation we can say that $\mathrm{tr}\left\{\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2\boldsymbol{A}_1\right\} = F > 0$ and $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{B}_2\boldsymbol{A}_1\boldsymbol{B}_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = H > 0$ and so we can say,

$$\begin{aligned}\mathrm{Var}\left\{\delta_T(\widehat{\boldsymbol{\mu}}_k)\right\} =& \frac{1}{2T^{2\alpha}}[N + F + 2T^{\alpha}H + T^{\alpha}C + 4D],\\ =& \frac{2H+C}{2T^{\alpha}} + \frac{N+F+D}{2T^{2\alpha}}.\end{aligned}$$

Clearly as $T\to\infty$ then $\mathrm{Var}\left\{\delta_T(\widehat{\boldsymbol{\mu}}_k)\right\}\to 0$ and so condition **A2** is satisfied. Since both conditions are now satisfied we have established that $\Pr(\delta_T(\widehat{\boldsymbol{\mu}}_k)\leq 0|C(k)=1)\to 0$ as $T\to\infty$.

■

B.2 Proof of Proposition 4.2

We now consider the case of the distance between classes diverging, i.e. $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|\to\infty$ for a fixed T. Here we define $|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2| = \sqrt{\sum_{i=1}^{N}|(\boldsymbol{\mu}_1)_i-(\boldsymbol{\mu}_2)_i|}$. To this end we define a different scaling of the divergence criterion,

$$\begin{aligned}\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k) =& \frac{\Delta(\widehat{\boldsymbol{\mu}}_k)}{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2},\\ =& \frac{1}{2|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2}\left\{+(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_2) - (\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_1) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right\}.\end{aligned}$$

Following a similar logic to the proof of Proposition 4.1 we aim to show that in the limit $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ the probability of misclassification, $\Pr(D(\widehat{\boldsymbol{\mu}}_k) = 2|C(k) = 1)$ will tend to 0. This is equivalent to showing that $\Pr(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k) \leq 0|C(k) = 1) \to 0$ in the same limit which immediately follows if we satisfy the following conditions in the limit as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$: **B1:** $E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] \to K$ where $K > 0$ and **B2:** $\text{Var}\left\{\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right\} \to 0$.

Expectation

We first consider the expected value of $\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)$. Using the results from the proof of Proposition 4.1 it is readily seen that,

$$E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{1}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^2}\left\{-N + \text{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right\}.$$

We assume that the terms N, $\text{tr}\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\}$ and $\log|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|$ do not depend upon the distance between classes and so we will replace these three terms by the constant Q. We now consider the third term in the bracket, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. First we rewrite $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in the form,

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|\,\boldsymbol{v},$$

where $\boldsymbol{v} = [v_1, \ldots, v_N]'$ a vector of constants. The third term can then be rewritten,

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2\,\boldsymbol{v}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{v} = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2\,R,$$

where R is a positive constant due to $\boldsymbol{\Sigma}_2$, and therefore $\boldsymbol{\Sigma}_2^{-1}$ being positive definite. Putting these terms into the expectation we get,

$$E\left[\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right] = \frac{Q}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^2} + \frac{R}{2}.$$

Clearly as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ then $E\left[\delta(\widehat{\boldsymbol{\mu}}_k)\right] \to \frac{R}{2}$ which is a positive constant thus condition **B1** is satisfied.

Variance

We now consider the variance of $\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)$. Using the results from Section B.1 we can say that,

$$\begin{aligned}\text{Var}\left\{\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right\} =& \frac{1}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^4}\left[N + \text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} + 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right. \\ &\left. +(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + 2\text{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}\right].\end{aligned}$$

We first look at the terms which do not depend on $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|$ namely N, $\text{tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\}$ and $\text{tr}\left\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right\}$. These terms can again be collected into one constant term, U. We have already stated that the fourth term in the brackets can be written as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2 R$. The third term can also be re written,

$$2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2 \boldsymbol{v}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{v} = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|^2 V.$$

Putting these terms into the variance we get,

$$\text{Var}\left\{\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right\} = \frac{U}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^4} + \frac{R+V}{2\left|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right|^2}.$$

Clearly in the limit $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$ then $\text{Var}\left\{\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k)\right\} \to 0$ and so the condition **B2** is satisfied. We have therefore satisfied both conditions for this proof and have established that $\Pr(\delta_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_k) \leq 0 | C(k) = 1) \to 0$ as $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \to \infty$.

■

Appendix C

Proofs for Chapter 5

C.1 Proof of Proposition 5.1

We consider the variance of the elements of the estimated LWS matrix. The estimate of the (p, q) element of the LWS at level j and location k is denoted as $\widehat{S}_{j,k}^{(p,q)} = \sum_{l=1}^{J} A_{jl}^{-1} \tilde{I}_{j,k}^{(p,q)}$, where $\tilde{I}_{j,k}^{(p,q)}$ is the smoothed periodogram and A_{jl} is the autocorrelation wavelet inner product matrix. The variance of this estimate is therefore,

$$\begin{aligned} \mathrm{Var}\left\{\widehat{S}_{j,k}^{(p,q)}\right\} &= \mathrm{Var}\left\{\sum_{l=1}^{J} A_{jl}^{-1} \tilde{I}_{j,k}^{(p,q)}\right\}, \\ &= \sum_{l_1=1}^{J}\sum_{l_2=1}^{J} A_{jl_1}^{-1} A_{jl_2}^{-1} \,\mathrm{cov}\left(\tilde{I}_{l_1,k}^{(p,q)}, \tilde{I}_{l_2,k}^{(p,q)}\right). \end{aligned}$$

Smoothing of the periodogram is performed using a box kernel with smoothing parameter M. The smoothed periodogram therefore has the form $\tilde{I}_{l,k}^{(p,q)} = (2M + 1)^{-1} \sum_{m=k-M}^{k+M} I_{l,m}^{(p,q)}$. Consequently the covariance between elements of the smoothed

periodogram is given by,

$$\text{cov}\left(\tilde{I}^{(p,q)}_{l_1,k}, \tilde{I}^{(p,q)}_{l_2,k}\right) = \frac{1}{(2M+1)^2} \sum_{m_1=k-M}^{k+M} \sum_{m_2=k-M}^{k+M} \text{cov}\left(I^{(p,q)}_{l_1,m_1}, I^{(p,q)}_{l_2,m_2}\right).$$

The remaining challenge is to calculate the covariance between elements of the raw periodogram. To find this we use the definition of the raw periodogram, $I^{(p,q)}_{j,k} = d^{(p)}_{j,k} d^{(q)}_{j,k}$, where $d^{(p)}_{j,k}$ is the empirical wavelet coefficient for channel p at level j and location k.

$$\text{cov}\left(I^{(p,q)}_{j,k}, I^{(p,q)}_{l,m}\right) = E\left[I^{(p,q)}_{j,k} I^{(p,q)}_{l,m}\right] - E\left[I^{(p,q)}_{j,k}\right] E\left[I^{(p,q)}_{l,m}\right].$$

As established in Park et al. (2014) the expected value of the raw periodogram is $E\left[I^{(p,q)}_{j,k}\right] = \sum_l A_{jl} S^{(p,q)}_l(k/T) + \mathcal{O}(T^{-1})$. We therefore focus our attention on the first term. Using the results from Isserlis (1918) we can rewrite this term as,

$$\begin{aligned} E\left[I^{(p,q)}_{j,k} I^{(p,q)}_{l,m}\right] =& E\left[d^{(p)}_{j,k} d^{(q)}_{j,k} d^{(p)}_{l,m} d^{(q)}_{l,m}\right], \\ =& E\left[d^{(p)}_{j,k} d^{(q)}_{j,k}\right] E\left[d^{(p)}_{l,m} d^{(q)}_{l,m}\right] + E\left[d^{(p)}_{j,k} d^{(p)}_{l,m}\right] E\left[d^{(q)}_{j,k} d^{(q)}_{l,m}\right] + E\left[d^{(p)}_{j,k} d^{(q)}_{l,m}\right] E\left[d^{(q)}_{j,k} d^{(p)}_{l,m}\right]. \end{aligned}$$

We now concentrate on the term, $E\left[d^{(p)}_{j,k} d^{(q)}_{l,m}\right]$. Using the definition of the empirical wavelet coefficients, $d^{(p)}_{j,k} = \sum_t X^{(p)}_t \psi_{j,k}(t)$, where $X^{(p)}_t = \sum_l \sum_m \sum_r V^{(p,r)}_l(m/T) \psi_{l,m}(t) z^{(r)}_{l,m}$, and the covariance property of the random elements, $\text{cov}\left(z^{(r)}_{jk}, z^{(r')}_{j'k'}\right) = \delta_{rr'}\delta_{jj'}\delta_{kk'}$, we can express this expectation as follows,

$$\begin{aligned} E\left[d^{(p)}_{j,k} d^{(q)}_{l,m}\right] &= \sum_{h=1}^{J} \sum_n \sum_{r=1}^{\min(p,q)} V^{(p,r)}_h(n/T) V^{(q,r)}_h(n/T) \sum_t \psi_{h,n}(t)\psi_{j,k}(t) \sum_{t'} \psi_{h,n}(t')\psi_{l,m}(t'), \\ &= \sum_{h=1}^{J} \sum_n S^{(p,q)}_h(n/T) \sum_t \psi_{h,n}(t)\psi_{j,k}(t) \sum_{t'} \psi_{h,n}(t')\psi_{l,m}(t'), \\ &= \sum_{h=1}^{J} S^{(p,q)}_h\left(\frac{k+m}{2T}\right) \sum_n \sum_t \psi_{h,n}(t)\psi_{j,k}(t) \sum_{t'} \psi_{h,n}(t')\psi_{l,m}(t') + \mathcal{O}(T^{-1}). \end{aligned}$$

Here the final line uses the Lipschitz continuous property of the LWS as well as the fact that$\sum_t \psi_{h,n}(t)\psi_{j,k}(t)$ is also finite. Rearranging the last terms and using the substitution $\tau = t' - t$ gives,

$$\begin{aligned}
E\left[d_{j,k}^{(p)} d_{l,m}^{(q)}\right] &= \sum_{h=1}^{J} S_h^{(p,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \sum_{t} \psi_{j,k-t}\psi_{l,m-t-\tau} \sum_{n} \psi_{h,n-t}\psi_{h,n-t-\tau} + \mathcal{O}(T^{-1}), \\
&= \sum_{h=1}^{J} S_h^{(p,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \sum_{t} \psi_{j,k-t}\psi_{l,m-t-\tau} \Psi_h(\tau) + \mathcal{O}(T^{-1}), \\
&= \sum_{h=1}^{J} S_h^{(p,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \sum_{t} \psi_{j,k-t}\psi_{l,k-t+m-k-\tau} \Psi_h(\tau) + \mathcal{O}(T^{-1}), \\
&= \sum_{h=1}^{J} S_h^{(p,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) + \mathcal{O}(T^{-1}).
\end{aligned}$$

Where $\Psi_j(\tau) = \sum_k \psi_{j,k}\psi_{j,k+\tau}$ is the autocorrelation wavelet defined in Nason et al. (2000) and $\Psi_{jl}(\tau) = \sum_k \psi_{j,k}\psi_{l,k+\tau}$ is defined in Fryzlewicz and Nason (2006). Putting this into the expression for $E\left[I_{j,k}^{(p,q)} I_{l,m}^{(p,q)}\right]$ gives,

$$\begin{aligned}
E\left[I_{j,k}^{(p,q)} I_{l,m}^{(p,q)}\right] = &\sum_{h} A_{jh} S_h^{(p,q)}(k/T) \sum_{h'} A_{lh'} S_{h'}^{(p,q)}(m/T) \\
&+ \sum_{h=1}^{J} S_h^{(p,p)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) \\
&\times \sum_{h'=1}^{J} S_{h'}^{(q,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_{h'}(\tau) \\
&+ \sum_{h=1}^{J} S_h^{(p,p)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) \\
&\times \sum_{h'=1}^{J} S_{h'}^{(q,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_{h'}(\tau) + \mathcal{O}(T^{-1}).
\end{aligned}$$

Putting this into the expression for the covariance gives,

$$\begin{aligned}\operatorname{cov}\left(I_{j,k}^{(p,q)}, I_{l,m}^{(p,q)}\right) = &\sum_{h=1}^{J} S_h^{(p,p)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) \\ &\times \sum_{h'=1}^{J} S_{h'}^{(q,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) \\ &+ \sum_{h=1}^{J} S_h^{(p,p)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_h(\tau) \\ &\times \sum_{h'=1}^{J} S_{h'}^{(q,q)} \left(\frac{k+m}{2T}\right) \sum_{\tau} \Psi_{jl}(k-m+\tau)\Psi_{h'}(\tau) + \mathcal{O}(T^{-1}).\end{aligned}$$

■

Bibliography

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*. Wiley, 2nd edition.

Böhm, H., Ombao, H. C., von Sachs, R., and Sanes, J. (2010). Classification of multivariate non-stationary signals: The SLEX-shrinkage approach. *Journal of Statistical Planning and Inference*, 140(12):3754–3763.

Brockwell, P. and Davis, R. (2009). *Time series: theory and methods*. Springer.

Caiado, J., Crato, N., and Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10):2668–2684.

Cappé, O. (2002). A Bayesian approach for simultaneous segmentation and classification of count data. *Signal Processing, IEEE Transactions on*, 50(2):400–410.

Cappé, O., Moulines, E., and Ryden, T. (2006). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer.

Cardinali, A. and Nason, G. (2010). Costationarity of locally stationary time series. *Journal of Time Series Econometrics*, 2(2):1–35.

Cardinali, A. and Nason, G. P. (2008). Locally stationary wavelet packet processes : nonstationarity detection and model fitting . *University of Bristol Technical Report*.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction*. Chapman and Hall, sixth edition.

Cho, H. and Fryzlewicz, P. (2014). Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. Technical report, The London School of Economics and Political Science.

Cohen, E. and Walden, A. (2010). A statistical study of temporally smoothed wavelet coherence. *IEEE Transactions on Signal Processing*, 58(6):2964–2973.

Cohen, E. and Walden, A. (2011). Wavelet coherence for certain nonstationary bivariate processes. *IEEE Transactions on Signal Processing*, 59(6):2522–2531.

Cohen, L. (1989). Time-frequency distributions - a review. *Proceedings of the IEEE*, 77:941–981.

Coifman, R. and Wickerhauser, M. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718.

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

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

Dahlhaus, R. (2000b). Graphical interaction models for multivariate time series1. *Metrika*, 51(2):157–172.

Dahlhaus, R. (2012). Locally stationary processes. *Handbook of Statistics*.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

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.

Eckley, I. and Nason, G. (2005). Efficient computation of the discrete autocorrelation wavelet inner product matrix. *Statistics and Computing*, 15(2):83–92.

Fiecas, M. and Ombao, H. C. (2011). The generalized shrinkage estimator for the analysis of functional connectivity of brain signals. *The Annals of Applied Statistics*, 5(2):1102–1125.

Fiecas, M., Ombao, H. C., and Linkletter, C. (2010). Functional connectivity: Shrinkage estimation and randomization test. *NeuroImage*, 49(4):3005–14.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Fryzlewicz, P. (2008). A waveletFisz approach to spectrum estimation. *Journal of Time Series Analysis*, 29(5):868–880.

Fryzlewicz, P. and Nason, G. (2006). HaarFisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society. Series B*, 68(4):611–634.

Fryzlewicz, P. and Ombao, H. C. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *Journal of the American Statistical Association*, 104(485):299–312.

Fryzlewicz, P., van Bellegem, S., and Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Annals of the Institute of Statistical Mathematics*, 55(4):737–764.

Gombay, E. (2008). Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3):451–464.

Huang, H.-Y., Ombao, H. C., and Stoffer, D. S. (2004). Discrimination and classification of nonstationary time series using the SLEX model. *Journal of the American Statistical Association*, 99(467):763–774.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a

normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139.

Jolliffe, I. T. (1972). Discarding Variables in a Principle Component Analysis. *Journal of the Royal Statistical Society. Series C*, 22(1):21–31.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association*, 93(441):328–340.

Kayhan, A., El-Jaroudi, A., and Chaparro, L. F. (1994). Evolutionary periodogram for nonstationary signals. *IEEE Transactions on Signal Processing*, 42(6):1527–1536.

Killick, R., Eckley, I. A., and Jonathan, P. (2013). A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electronic Journal of Statistics*, 7:1167–1183.

Koopmans, L. H. (1964). On the multivariate analysis of weakly stationary stochastic processes. *The Annals of Mathematical Statistics*, 35:1765 – 1780.

Koopmans, L. H. (1975). *The Spectral Analysis of Time Series*. Academic Press.

Krzemieniewska, K., Eckley, I., and Fearnhead, P. (2014). Classification of nonstationary time series. *Stat*, 3(1):144–157.

Krzemieniewska, K. I. (2013). *Classification of non-stationary time series*. PhD thesis, Lancaster University.

Kumar, A. and Fuhrmann, D. (1992). A new transform for time-frequency analysis. *IEEE Transactions on Signal Processing*, 40(7):1697–1707.

Lee, T. (1997). A simple span selector for periodogram smoothing. *Biometrika*, 84(4):965–969.

Liu, S. and Maharaj, E. A. (2013). A hypothesis test using bias-adjusted AR estimators for classifying time series in small samples. *Computational Statistics & Data Analysis*, 60:32–49.

MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete- valued Time Series*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Mallat, S. (1989a). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.

Mallat, S. (1989b). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of the American mathematical society*, 315(1):69–87.

McCabe, G. (1984). Principal variables. *Technometrics*, 26(2):137–144.

Meenakshi, A. and Rajian, C. (1999). On a product of positive semidefinite matrices. *Linear algebra and its applications*, 295:9–12.

Meyer, Y. (1986). Principe d'incertitude, bases Hilbertiennes et algebres d'operatuers. *Seminaire N. Bourbaki*, 662:209–223.

Nam, C. F. H., Aston, J. A. D., Eckley, I. A., and Killick, R. (2014). The Uncertainty of Storm Season Changes: Quantifying the Uncertainty of Autocovariance Changepoints. *Technometrics*.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. Springer.

Nason, G. (2013). A test for second order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society: Series B*, 75(5):879–904.

Nason, G. and Silverman, B. (1995). The stationary wavelet transform and some statistical applications. In Antoniadis, A. and Oppenheim, G., editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 281–299. Springer New York.

Nason, G., von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society. Series B*, 62(2):271–292.

Ombao, H., von Sachs, R., and Guo, W. (2000). Estimation and inference for time-varying spectra of locally stationary SLEX processes In Memory of Jonathan A. Raz. *Proceedings of the 2nd International Symposium on Frontiers of Time Series Modeling*.

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

Ombao, H. C., Raz, J. A., von Sachs, R., and Guo, W. (2002). The SLEX Model of a Non-Statioary Random Process. *Ann. Inst. Statist. Math.*, 54(1):171–200.

Ombao, H. C. and Van Bellegem, S. (2008). Evolutionary coherence of nonstationary signals. *IEEE Transactions on Signal Processing*, 56(6):2259–2266.

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

Park, T., Eckley, I., and Ombao, H. (2014). Estimating Time-Evolving Partial Coherence Between Signals via Multivariate Locally Stationary Wavelet Processes. *IEEE Transactions on Signal Processing*, 62(20):5240–5250.

Priestley, M. (1988). *Non-linear and non-stationary time series analysis*. Academic Press.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):204–237.

Priestley, M. B. (1981a). *Spectral Analysis and Time Series*, volume 1. Academic Press.

Priestley, M. B. (1981b). *Spectral Analysis and Time Series*, volume 2. Academic Press.

Sakiyama, K. and Taniguchi, M. (2004). Discriminant analysis for locally stationary processes. *Journal of Multivariate Analysis*, 90(2):282–300.

Sanderson, J., Fryzlewicz, P., and Jones, M. (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, 97(2):435–446.

Schuster, E. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *The Annals of Mathematical Statistics*, 43(1):84–88.

Scott, A. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512.

Shumway, R. (1982). Discriminant analysis for time series. In Krishnaiah, P. and Kanal, L., editors, *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2 of *Handbook of Statistics*, pages 1 – 46. Elsevier.

Shumway, R. and Stoffer, D. (2000). *Time series analysis and its applications*. Springer, second edition.

Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis. *Statistics & Probability Letters*, 63(3):307–314.

Slutsky, E. (1925). Über stochastische asymptoten und grenzwerte. *Metron*, 5:3–89.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley and Sons, Inc.

von Sachs, R. and Neumann, M. H. (2000). A Wavelet-Based Test for Stationarity. *Journal of Time Series Analysis*, 21(5):597–613.

Walden, A. and Cohen, E. (2012). Statistical Properties for Coherence Estimation From Evolutionary Spectra. *IEEE Transactions on Signal Processing*, 60(9):4586–4597.

Wickerhauser, M. V. (1994). *Adapted Wavelet Analysis: From Theory to Software*. AK Peters Series. Taylor & Francis.