Modelling and Estimation of Time Series with Long Memory

Keerati Suibkitwanchai, B.Eng., M.Sc (Dist.)

Submitted for the degree of Doctor of

Philosophy at Lancaster University.

July 2022

Abstract

Many real-world time series have been observed to have strong positive correlation between their long-term observed values, and this behaviour is known as long memory or long-range dependence. However, many statistical models and estimation techniques are built under the assumption of short memory (or sometimes complete independence) and identically distributed data. This challenges the modelling and estimation of time series with long memory. In the first half of this thesis, we investigate several parametric models for time series with long memory, which are commonly known as fractional models. We focus on the challenge of parameter estimation from sampled time series, and compare numerous existing and novel methods in wide-ranging simulation studies. The estimation results from all these methods are provided and compared among each other to argue that a novel method, known as the debiased Whittle likelihood estimator, originally proposed for time series with short memory, is also the most appropriate method for long memory in terms of mean squared error, consistency, asymptotic efficiency, and computational cost. We then implement this estimator to study long-memory behaviour found in real-world financial time series, as an example. In the second half of this thesis, we investigate two other applications

of time series with long-memory behaviour in health sciences and sport sciences. Both applications are concerned with high-frequency tracking of the movement of individuals, the former concerned with monitoring activity levels of patients with advanced dementia, and the latter concerned with footballers' movements during professional matches. In both applications, the observed time series exhibit cycles, trends, and changes in variability, as well as long-memory behaviour. Therefore parametric models are difficult to construct for the time series of these applications, and we instead propose nonparametric measures providing summary statistics jointly capturing long-memory behaviour alongside other summary statistics of interest.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr Adam Sykulski, who has given his full support to my PhD studies. His advice made me feel more confident in conducting my research and writing this thesis. His comments and feedback on my works were valuable for me to improve my research and other important skills for my PhD studies. He is a great supervisor with whom I can discuss any problem, whether or not it is related to my academic work for this thesis.

I also would like to thank all academic professors, professional and support staff, colleagues, and friends who supported and helped me develop my work in this thesis. These include all co-authors of my first published paper, especially Dr Guillermo Perez Algorta, an academic professor in the Division of Health Research at Lancaster University, who has supervised and helped me a lot regarding the collection of data and their analysis on health sciences for this paper. Another project in sport sciences in this thesis has been conducted with a lot of help from Dr Ian Cowling, a chief technology officer of Sportlight Technology Ltd., who has provided the data for our analysis and given me a lot of useful comments for my statistical analysis of these data in sport sciences.

Lastly, I am grateful to my parents for their support in everything that I have made throughout my PhD studies.

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.

Keerati Suibkitwanchai

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Acceleration
ACVF	Autocovariance Function
ACVS	Autocovariance Sequence
ANOVA	Analysis of Variance
AR	Autoregressive
ARFIMA	Autoregressive Fractionally Integrated Moving Average
CB	Centre-Backs
CM	Central Midfielders
CPU	Central Processing Unit
DEC	Deceleration
DFA	Detrended Fluctuation Analysis
DWLE	Debiased Whittle Likelihood Estimator
ENMO	Euclidean Norm Minus One
FAST	Functional Assessment Staging Tool
FB	Full-Backs
FBM	Fractional Brownian Motion

FD	Fractionally Differenced
FFT	Fast Fourier Transform
FGN	Fractional Gaussian Noise
FW	Forwards
GK	Goalkeeper
GPS	Global Positioning Systems
HS	High Speed
IS	Interdaily Stability
IV	Intradaily Variability
LIDAR	Light Detection and Ranging
LPRE	Log-Periodogram Regression Estimator
MA	Moving Average
MAE	Mean Absolute Error
MCS	Multiple-Camera Systems
MFDFA	Multifractal Detrended Fluctuation Analysis
MLE	Maximum Likelihood Estimator
PoV	Proportion of Variance
PPL	Pure Power Law
PSD	Power Spectral Density
RMSD	Root-Mean-Square Deviation
SD	Standard Deviation
SP	Sprint

VHS Video Home Systems

WLE Whittle Likelihood Estimator

WM Wide Midfielders

Chapter 1

Introduction to Long-Memory Processes

Time series analysis has been used to explain observed time-dependent data in various aspects. A widely used way of representing this time-dependence is via the autocovariance, or its normalisation by the variance known as the autocorrelation. These functions measure the covariance and correlation of values of time series at two different points of time, given by t_1 and t_2, where their difference is called lag $\tau = t_2 - t_1$. When a time series is second-order stationary such that its first two central moments do not depend on time, both autocovariance and autocorrelation at a particular lag τ are constant for any times t_1 and t_2 that are lag τ apart. However, the autocovariance and the autocorrelation of a nonstationary time series, which is referred to as any time series without second-order stationarity in this thesis, depend on both lag τ and individual points of time. It is usually the case that both functions, the autocovariance and the autocorrelation, generally decrease over lag τ. If these func-

tions drop rapidly, like an exponential decay, or reach zero over a small number of lags τ, the time series exhibits *short-memory* behaviour. In contrast, a more gradual decrease of these functions over a large number of lags τ, like a power decay, represents *long-memory* behaviour of the time series. This behaviour corresponds to a strong positive autocorrelation between values at distant points of time series, and this is referred to as *long-range dependence*. More technical aspects of long memory are deferred to Chapter 2. Time series in many real-world applications exhibit long-memory behaviour with slow decay of its autocovariance, which has been empirically observed.

The methodological treatment of time series with long memory is more challenging than short memory due to the high correlation coefficients found in the autocorrelation, even at large lags. Many statistical models and their estimation techniques are classically built from the assumptions of independence and identically distributed data, or from short-memory time series, and do not trivially extend to time series with long memory as shall become apparent. In this thesis, we will explore the modelling and estimation of parameters that quantify the degree of long memory or long-range dependence in time series. Some of the modelling and estimation techniques will come directly from the autocovariance or autocorrelation, but others will use alternative representations such as the power spectral density function in the frequency domain (also referred to as the spectrum in this thesis), or by directly modelling and estimating the long-term variability or volatility of time series. Both existing and novel estimation methods for long-memory parameters will be investigated. Related statistical measures, which capture other properties such as cyclical behaviour,

will also be evaluated and compared alongside the estimation results of long-memory parameters. In particular, in our application chapters, we will investigate the joint estimation of multiple parameters and summary statistics, to investigate the ability of capturing long memory in real-world datasets that are also influenced by other properties such as cycles, trends, and changes of variability.

1.1 Long-memory Parameters

Long-memory parameters have been studied in many research works. Hurst (1951) introduced a statistical measure called the rescaled range, which is the range of the cumulative sum of a time series (after the mean has been subtracted), divided by the standard deviation, and this has been often referred to as R/S analysis. Hurst (1951) found that there is a strong linear relationship between the rescaled range from n observed values and the logarithm of n, where n is a value between a small positive number and the total length of the time series. A parameter representing the estimated slope from the linear regression line of this relationship is nowadays known as the Hurst exponent (H). Hurst (1951) also found that the Hurst exponent can be estimated from the slope of the logarithm of the spectrum of the time series due to its power-law relationship. The Hurst exponent can be used to measure the degree of long-range dependence. The long-term positive autocorrelation of a long-memory process results in a higher value of the Hurst exponent than a short-memory process. An example of a process that has the Hurst exponent as its model parameter is fractional Brownian motion (FBM) defined in terms of a stochastic integral equation (Mandel-

brot and Van Ness, 1968). FBM is a nonstationary process which is a generalisation of Brownian motion, which is classified as a Lévy process. The difference between FBM and Brownian motion is that the increment, or the first-order difference process, of the former allows for the dependence or correlation of their observed values, while the increment of the latter is always Gaussian white noise with uncorrelated observations. Thus, the increment of FBM is called fractional Gaussian noise (FGN). This process is stationary and also has the Hurst exponent as its model parameter.

Another long-memory parameter found in the literature is the order of differencing "d", which is used in the fractionally differenced (FD) process proposed by Granger and Joyeux (1980) and Hosking (1981). It is a special case of the autoregressive fractionally integrated moving average or ARFIMA(p, d, q) process with the zeroth-order of autoregressive and moving average parts ($p = 0$ and $q = 0$, respectively), and therefore the FD process can be referred to as an ARFIMA$(0, d, 0)$ process. We note that sometimes FARIMA is used instead of ARFIMA in the literature. In general, the value of d for the FD process can be any real number. Another process using a long-memory parameter similar to H and d is a pure power law (PPL) process with parameter γ. This parameter is used to explain the power-law behaviour of the spectrum of the process at all observed frequencies. The PPL process is similar to other long-memory processes such as FBM and the FD process, at low frequencies, as we shall study further in Chapter 2. There are also widely used nonparametric approaches for capturing long memory, for example, Peng et al. (1994) introduced a long-memory summary statistic describing the scale invariance (or self-similarity) of the fluctuation of time series, and named a statistical procedure to calculate this

statistic as detrended fluctuation analysis (DFA). The summary statistic is called the scaling exponent and defined as α, and is related to H as we shall show in Chapter 2.

Among all long-memory parameters and summary statistics mentioned above, we will mainly focus on the Hurst exponent (H) and the scaling exponent from detrended fluctuation analysis (α) in this thesis. The estimation of H from several parametric methods and the nonparametric calculation of α will be explored in detail.

1.2 Estimators

Parameters of each fractional process (FBM, FGN, FD, or PPL) can be estimated with several parametric methods. Many properties are considered as criteria to select the best estimator among them. These properties include unbiasedness, low measurement errors, consistency, efficiency, and low computational cost. Fox and Taqqu (1986) investigated the maximum likelihood estimator (MLE) for the estimation of long-memory parameters. The advantage of this method is the consistency and asymptotic efficiency of the estimator with a convergence rate of $1/\sqrt{n}$, where n is the length of the fractional process. This is also called $\sqrt{n}$-consistency. However, the log-likelihood function of this estimator (see Brockwell (1987)) consists of the inversion of the autocovariance matrix and this contributes to very high computational cost for processes with high n, which is at least $\mathcal{O}(n^2)$ in general. We therefore also consider an estimator with lower computational cost, and this is an approximated version of the MLE in the frequency domain called the Whittle likelihood estimator (WLE) (Whittle, 1953). The log-likelihood function of this estimator utilises the

power spectral density function and the periodogram, which is the asymptotically unbiased estimator of the power spectral density function itself, and it still maintains the consistency and asymptotic efficiency of the estimate with the same convergence rate as the maximum likelihood estimator assuming certain assumptions are satisfied (Dzhaparidze and Kotz, 1986). However, the periodogram has been shown to have large bias for certain processes with finite samples, and this bias can cause substantial bias of estimates from the WLE (Contreras-Cristán et al., 2006). To alleviate this problem, Sykulski et al. (2019) proposed its debiased version with the replacement of the power spectral density function by the expected periodogram. However, they only proved consistency and asymptotic efficiency with some short-memory processes such as the Matérn and autoregressive (AR) processes in their theoretical and simulation studies. It is therefore of interest to see if the debiased Whittle likelihood estimator (DWLE) can be applied to long-memory processes, and we shall investigate this idea in this thesis.

1.3 Applications

The measurement of long memory has been conducted with time series of many applications in various fields of study, for instance, computer communications (Beran et al., 1995; Paxson, 1997), ecology (Wang et al., 2011), thermodynamics (Doucoure et al., 2016), hydrology (Hurst, 1951; Molz and Liu, 1997), oceanography (Sanderson et al., 1990; Lilly et al., 2017), atmospheric pollutants (Knight et al., 2017), and wind energy (Knight and Nunes, 2019). In this thesis, we will perform an analysis of long-

memory time series derived from financial data. In fact, financial time series are a widely-studied application area of long-memory behaviour as we shall now discuss.

1.3.1 Financial Time Series

In finance, the measurement of long memory has been performed with either the estimation of long-memory parameters or the modelling of fractional processes fitted to the observed data. Couillard and Davison (2005) studied the log returns of S&P 500 index and stock prices from three large capitalisation companies, and reported them as nonstationary Brownian motions according to the estimated Hurst exponents from the adjusted R/S analysis introduced by Annis and Lloyd (1976). Qian and Rasheed (2007) measured the Hurst exponent from the daily return of Dow Jones index with the standard R/S analysis (Hurst, 1951). They suggested that its time series is not totally random noise as the long-memory behaviour can be detected according to the high local Hurst exponent changing over the observed time periods. Corazza and Malliaris (2002) used a modified R/S test from Lo (1991) to check whether the daily returns of foreign exchange rates correspond to Brownian motion. They found that the local Hurst exponents are consistent with persistent long-range dependence of its increment process. These research works show examples of how long memory-behaviour has been detected in various financial assets, but to varying degrees, and motivates the need for good models and estimation methods to capture these differences. Apart from multiple versions of R/S analysis, some other methods that have been used to estimate the Hurst exponent from financial time series include multiresolution analysis (Mallat, 1989; Karuppiah and Los, 2005; Kyaw et al., 2006),

and the generalised Hurst exponent approach (Di Matteo et al., 2003; Barunik and Kristoufek, 2010; Sensoy, 2013). Some research works suggested to use the FD or ARFIMA$(0, d, 0)$ process for modelling the time series of financial assets (Ding et al., 1993; Man, 2003; Ellis and Wilson, 2004; Xiu and Jin, 2007). However, in this thesis, we will instead provide an example of the estimation of the Hurst exponent (H) from financial data, as the order of differencing (d) of the FD process can be theoretically predicted from the Hurst exponent by using a single relationship, which will be explained later. The financial data we shall study are log-volatility time series of four stock indices. Gatheral et al. (2018) also studied these time series with different financial assets and found that they behave like nonstationary fractional Brownian motion exhibiting long-memory behaviour. Unlike this work, we will estimate both global and time-local Hurst exponents from log-volatility time series along with their uncertainty intervals from our own estimation methods. It is of interest to compare how well our methods estimate the Hurst exponent from real financial data that have already been analysed by different estimators in Gatheral et al. (2018).

In the second half of this thesis, high-frequency time series data from two other applications in health sciences and sport sciences will be investigated in detail. Unlike financial data, we will measure their long-memory behaviour by the scaling exponent (α) from DFA and related nonparametric summary statistics as the data are more complicated with trends and cycles, and therefore parametric models do not work well here. Related previous works of estimating long-memory behaviour in both areas will be provided later in Chapters 5 and 6.

1.4 Contributions

This thesis is divided into three works with five chapters excluding the introduction and conclusion. Our main work on the estimation and modelling of fractional processes with long-memory parameters is explained in Chapters 2–4. The parametric estimation will be applied to simulated time series and log-volatility time series from financial assets. Then in Chapters 5–6, we will conduct two other research studies on the applications of nonparametric statistical measures related to long memory to high-frequency data. The first research has been published in the PLOS ONE Journal with the title of "*Nonparametric Time Series Summary Statistics for High-Frequency Accelerometry Data from Individuals with Advanced Dementia*," cited as Suibkitwanchai et al. (2020), and this paper is given in Chapter 5. The second research uses some nonparametric measures from this published paper along with new measures to study the impact of fatigue on performance of footballers, and this is explained in Chapter 6. All works in this thesis are concluded in the final chapter or Chapter 7. A brief outline of our works in Chapters 2–6 is given as follows:

Chapter 2: Background reviews the properties of four fractional processes studied in this thesis. The first two processes are nonstationary fractional Brownian motion (FBM) and its first-order difference process called stationary fractional Gaussian noise (FGN). Both depend on the Hurst exponent (H) with a value between zero and one, which determines their degree of long memory or long-range dependence. Another fractional process is the fractionally differenced (FD) process, which is a discrete-time fractional process depending on the order of differencing (d) of Gaus-

sian white noise. Lastly, the pure power law (PPL) process is another discrete-time fractional process used to model the power-law behaviour found in the spectrum of many time series with long memory. With each process, the conditions on the parameters for the existence of long memory and stationarity are also given.

Chapter 3: Estimation Methods describes existing and novel estimators for long-memory parameters. The maximum likelihood estimator is explained but not used in practice due to its high computational cost. Its approximated versions, the Whittle likelihood estimator and the debiased Whittle likelihood estimator, are used instead. The log-periodogram regression estimator is another method which estimates long-memory parameters by performing regression on the logarithm of the periodogram at low frequencies. We shall also discuss that for a finite-length nonstationary process, data tapering can be conducted before the estimation of the long-memory parameter to reduce bias from the spectral leakage of its spectrum.

Chapter 4: Simulation and Empirical Studies performs detailed simulation studies to assess the performance of the estimation methods provided in Chapter 3. In particular, we compare estimated values of the Hurst exponent from simulated time series of fractional Gaussian noise with various lengths and true values of the Hurst exponent. Quantification of the uncertainty of these estimates is also discussed. Lastly, we provide an example of our estimation of the Hurst exponent from the log-volatility time series of four financial assets reported in the Oxford-Man Institute's realized library (https://realized.oxford-man.ox.ac.uk).

Chapter 5: Nonparametric Statistics for High-Frequency Accelerometry Data from Individuals with Advanced Dementia provides our published

paper in the PLOS ONE Journal (Suibkitwanchai et al., 2020). The content of this chapter is exactly the same as the paper. The abstract of this paper is given as follows:

"Accelerometry data has been widely used to measure activity and the circadian rhythm of individuals across the health sciences, in particular with people with advanced dementia. Modern accelerometers can record continuous observations on a single individual for several days at a sampling frequency of the order of one hertz. Such rich and lengthy data sets provide new opportunities for statistical insight, but also pose challenges in selecting from a wide range of possible summary statistics, and how the calculation of such statistics should be optimally tuned and implemented. In this paper, we build on existing approaches, as well as propose new summary statistics, and detail how these should be implemented with high frequency accelerometry data. We test and validate our methods on an observed data set from 26 recordings from individuals with advanced dementia and 14 recordings from individuals without dementia. We study four metrics: Interdaily stability (IS), intradaily variability (IV), the scaling exponent from detrended fluctuation analysis (DFA), and a novel nonparametric estimator which we call the proportion of variance (PoV), which calculates the strength of the circadian rhythm using spectral density estimation. We perform a detailed analysis indicating how the time series should be optimally subsampled to calculate IV, and recommend a subsampling rate of approximately 5 minutes for the dataset that has been studied. In addition, we propose the use of the DFA scaling exponent separately for daytime and nighttime, to further separate effects between individuals. We compare the relationships between all these methods and show that they effectively capture different features of the time series."

We perform the analysis of high-frequency accelerometry data with the four statistical methods given in the abstract. All of them are nonparametric as no parametric assumptions are made for the distribution of their time series. Instead of the Hurst exponent from a parametric model, we calculate the scaling exponent from DFA to measure the degree of long memory in a time series. Several reasons are that (1) it is highly likely that the time series are non-Gaussian and non-stationary due to the existence of the cyclical trend of the circadian rhythm which makes a complete parametric model very difficult to construct, and (2) it would be complicated and require very high computational time to estimate parameters of long time series of high-frequency accelerometry data using any parametric estimation method in Chapter 3, due to the volume and length of time series considered. Apart from long-memory behaviour, we also measure the strength of circadian rhythm and the variability of time series from three other statistical methods: IS, IV, and PoV. Finally, we perform a comparison of numerical results from all methods applied to the time series of accelerometry data from two groups of people with advanced dementia, who received either normal treatment (non-intervention group) or special treatment (intervention group), and a group of people without dementia.

Chapter 6: The Study of Several Measures to Detect Fatigue in Sport Sciences is about the analysis of acceleration and related data from footballers to study fatigue due to their activity and movement during professional football matches. We received the data of their individual performance from Sportlight Technology Ltd. (https://www.sportlight.ai), and these data are divided into two datasets. The first set contains the information of position, speed, and acceleration of each player, while

the second set reports the events of *significant turns* defined as the sudden changes of players' movement from high deceleration to high acceleration in a short period of time. In this chapter, we report several measures including the total distance covered, the rate of change of distance covered at high acceleration in magnitude, and the total number of significant turns. Two nonparametric statistics from Chapter 5 including the intradaily variability (IV) and the scaling exponent from DFA are evaluated for the measurement of long memory in the time series of acceleration data from footballers. Because each acceleration time series is non-Gaussian and non-stationary, which is similar to what we found from the accelerometry time series in the previous chapter, it is better to use DFA than the parametric estimation methods proposed in Chapter 3, as once again a full parametric model is very difficult to construct for such complex time series.

The main goal of this chapter is to detect the difference of these measures and statistics from players in different positions, as well as the impact of fatigue to their performance as the game progresses across the 90 minutes.

Chapter 2

Background

In this chapter, theoretical concepts of continuous-time and discrete-time fractional processes along with their long-memory conditions are given. This chapter is divided into six sections. Section 2.1 explains some important functions and keywords related to time series analysis, and these will be used in the study of fractional processes in other sections. Section 2.2 provides a brief introduction of continuous-time nonstationary FBM and its Hurst exponent. The general form, the autocovariance function, and the spectrum of this fractional process are also defined. Section 2.3 shows these functions for discrete-time stationary FGN, which is the first-order increment or difference process of regularly-sampled FBM. The approximation of the spectrum that incorporates aliasing due to this sampling is mathematically given. Section 2.4 expresses the definition of the discrete-time FD process, and provides its several features in both the time and frequency domains. The FD process is either stationary or nonstationary depending on its parameter d representing the order of differencing of Gaussian white noise, which is explained in this section. Section 2.5 provides the

details of the discrete-time PPL process and the similarity between its spectrum and that of other fractional processes at low frequencies related to the power-law relationship. Lastly, Section 2.6 explains the requirements for long memory of a given time series process as derived from its autocovariance and its spectrum. These conditions are used to define the parameter regions for the existence of long memory from each fractional process studied in this chapter.

2.1 Keywords and Functions of Time Series

2.1.1 Stationarity

Stationarity is a fundamental concept used in time series analysis to determine the modelling of time series data or stochastic processes. Let $\{X_t \mid t \in T\}$ be a real-valued time series of length n generated from a stochastic process, where T is the domain of values that t is allowed to take. If T is a continuous domain (such as the set of real numbers $\mathbb{R}$) then we say the stochastic process is "continuous-time", conversely if T is a discrete domain (such as the set of integers $\mathbb{Z}$) then we say the stochastic process is "discrete-time". We shall work with both continuous- and discrete-time processes in this thesis, but keep our definitions general where possible. The time series $\{X_t \mid t \in T\}$ is strictly or strongly stationary if the unconditional joint probability distribution $F_X(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is identical to the unconditional joint probability distribution $F_X(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_k+\tau})$ such that $t_1, t_2, \ldots, t_k \in T$ and $t_1+\tau, t_2+\tau, \ldots, t_k+\tau \in T$ for any possible real value of lag τ and for any positive integer $k \leq n$. Furthermore, the time series is weakly or second-order stationary if

- The mean is finite and constant, i.e., $-\infty < \mathbb{E}[X_t] = \mu_X < \infty$.
- The variance is finite and constant, i.e., $-\infty < \mathrm{Var}[X_t] = \sigma_X^2 < \infty$.
- The autocovariance and autocorrelation only depend on lag τ.

The third assumption is about the property of autocovariance (or autocorrelation). This function is used to describe the covariance (or correlation) of a single time series at two different points of time such as t_1 and t_2. For a weakly stationary time series, the autocovariance solely depends on lag $\tau = t_2 - t_1$ such that for a given τ, the observed times t_1 and t_2 have no effect on the autocovariance (or autocorrelation). We provide the formal definitions in the next section.

In this thesis, we will use the word "stationary" as the same meaning as weakly stationary. Furthermore, the word "nonstationary" is used to represent the behaviour which is not weakly stationary such that at least one of these three given assumptions is not satisfied.

2.1.2 Autocovariance and Autocorrelation

The theoretical autocovariance of a stationary time series $\{X_t \mid t \in \mathbb{Z}\}$, is given at lag $\tau \in \mathbb{Z}$ by

$$\begin{aligned} s_{X,\tau} = \mathrm{Cov}[X_t, X_{t+\tau}] &= \mathbb{E}\left[(X_t - \mathbb{E}[X_t])(X_{t+\tau} - \mathbb{E}[X_{t+\tau}])\right] \\ &= \mathbb{E}[X_t X_{t+\tau}] - \mathbb{E}[X_t]\,\mathbb{E}[X_{t+\tau}] \\ &= \mathbb{E}[X_t X_{t+\tau}] - \mu_X^2 \\ &= s_{X,-\tau}, \end{aligned} \tag{2.1.1}$$

such that the autocovariance is symmetric in τ. The autocovariance at lag $\tau = 0$ is represented as $\mathbb{E}[X_t^2] - \mu_X^2$, which is equivalent to the second central moment or the variance of $\{X_t\}$, denoted as σ_X^2.

The definition above is given for a stationary time series sampled at integer times ($t \in \mathbb{Z}$), but applies to stationary stochastic processes in general. For example, if we consider a continuous-time stationary stochastic process ($t \in \mathbb{R}$), which we shall henceforth denote $\{X(t)\}$, then the definition of the autocovariance is identical to Equation (2.1.1), except replacing X_t by $X(t)$, and allowing $\tau \in \mathbb{R}$. We will use the term "autocovariance function" (ACVF) or $s_X(\tau)$ when considering a continuous-time stationary stochastic process, and the term "autocovariance sequence" (ACVS) or $s_{X,\tau}$ when considering a discrete-time stationary stochastic process or a sampled time series (from either a continuous or discrete-time stationary stochastic process). For a nonstationary process, the autocovariance and autocorrelation depend on both time t and lag τ.

In practice, we can calculate the sample autocovariance of the observed time series from $\{X_t \mid t \in T\}$ sampled uniformly at times $t = 1, 2, \ldots, n$ as

$$\hat{s}_{X,\tau} = \frac{1}{n} \sum_{t=1}^{n-\tau} (X_t - \bar{X})(X_{t+\tau} - \bar{X}), \qquad \tau = 0, \ldots, n-1, \tag{2.1.2}$$

where $\bar{X}$ is the sample mean of the observed time series from $\{X_t \mid t \in T\}$.

The autocorrelation is often used interchangeably with the autocovariance in many fields of study. However, in Statistics, these two terms are different such that the autocorrelation is in fact the normalisation of the autocovariance by the variance. This autocorrelation is often referred to as the autocorrelation coefficient, which is

defined as $\rho_{X,\tau}$ for a sampled time series or discrete-time process, and $\rho_X(\tau)$ for a continuous-time process. This coefficient always satisfies $|\rho_{X,\tau}| \leq 1$ or $|\rho_X(\tau)| \leq 1$ for any observable lag τ, and is expressed for discrete-time stationary processes by

$$\begin{aligned}
\rho_{X,\tau} = \mathrm{Cor}[X_t, X_{t+\tau}] &= \frac{\mathrm{Cov}[X_t, X_{t+\tau}]}{\sigma_{X_t}\sigma_{X_{t+\tau}}} \\
&= \frac{\mathbb{E}\left[(X_t - \mathbb{E}\left[X_t\right])(X_{t+\tau} - \mathbb{E}\left[X_{t+\tau}\right])\right]}{\sigma_{X_t}\sigma_{X_{t+\tau}}} \\
&= \frac{s_{X,\tau}}{\sigma_X^2} \\
&= \frac{s_{X,-\tau}}{\sigma_X^2} \\
&= \rho_{X,-\tau}, \qquad (2.1.3)
\end{aligned}$$

and similarly for continuous-time stationary processes replacing X_t with $X(t)$, and $\rho_{X,\tau}$ with $\rho_X(\tau)$. In practice, the sample autocorrelation of the observed time series from $\{X_t \mid t \in T\}$ with $t = 1, 2, \ldots, n$ is given by

$$\hat{\rho}_{X,\tau} = \frac{\sum_{t=1}^{n-\tau}(X_t - \bar{X})(X_{t+\tau} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}, \qquad \tau = 0, \ldots, n-1. \qquad (2.1.4)$$

The sample autocovariance (Equation (2.1.2)) and the sample autocorrelation (Equation (2.1.4)) are both asymptotically unbiased estimators of the autocovariance and the autocorrelation, respectively, assuming stationarity is satisfied.

2.1.3 Power Spectral Density Function

The power spectral density function, which is also referred to as the spectrum in this thesis, is used to show the distribution of the power of time series over frequency. Let $\{X(t)\}$ be a continuous-time and real-valued stationary process. The average power of this process is shown as an integral in either the time or frequency domain using

Parseval's theorem, expressed by

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |X(t)|^2 \, dt \; = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} |\mathscr{F}_\omega\{X(t)\}|^2 \, d\omega, \tag{2.1.5}$$

where $\mathscr{F}_\omega\{X(t)\} = \int_{-\infty}^{\infty} X(t)e^{-i\omega t}\, dt$ is the Fourier transform of $\{X(t)\}$ ($i \equiv \sqrt{-1}$ is the imaginary number). The integrand $|\mathscr{F}_\omega\{X(t)\}|^2$ is equivalent to the Fourier transform of the convolution of $\{X(t)\}$ with itself, and this is given by

$$|\mathscr{F}_\omega\{X(t)\}|^2 = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} X(t+\tau)X(t)\, dt \right] e^{-i\omega\tau}\, d\tau. \tag{2.1.6}$$

The power spectral density function of $\{X(t)\}$ at each frequency ω is defined as the integrand of its average power in the frequency domain, i.e.,

$$\tilde{S}_X(\omega) = \lim_{T\to\infty} \frac{1}{T} |\mathscr{F}_\omega\{X(t)\}|^2. \tag{2.1.7}$$

By substituting the Fourier transform from Equation (2.1.6) into this definition of the spectrum (Equation (2.1.7)), we can simplify the spectrum as

$$\tilde{S}_X(\omega) = \int_{-\infty}^{\infty} \mathbb{E}\left[X(t+\tau)X(t)\right] e^{-i\omega\tau}\, d\tau = \int_{-\infty}^{\infty} s_X(\tau) e^{-i\omega\tau}\, d\tau, \tag{2.1.8}$$

where $s_X(\tau)$ is the autocovariance function at lag τ of the process $\{X(t)\}$. From the above equation, the spectrum and the autocovariance function are Fourier transform pairs, i.e., $s_X(\tau) \leftrightarrow \tilde{S}_X(\omega)$. The inverse Fourier transform of the spectrum is given by

$$s_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{S}_X(\omega) e^{i\omega\tau}\, d\omega. \tag{2.1.9}$$

Similarly, for a discrete-time stationary process $\{X_t \mid t \in \mathbb{Z}\}$, the spectrum and the autocovariance sequence are related by the following equation:

$$S_X(\omega) = \sum_{\tau=-\infty}^{\infty} s_{X,\tau} e^{-i\omega\tau}, \tag{2.1.10}$$

where $s_{X,\tau}$ is defined in Equation (2.1.1) for $\tau \in \mathbb{Z}$. Hence, the spectrum is 2π-periodic such that $S_X(\omega) = S_X(\omega + 2\pi k)$ for $k \in \mathbb{Z}$. Its inverse Fourier transform is given by

$$s_{X,\tau} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_X(\omega) e^{i\omega\tau} \, d\omega. \tag{2.1.11}$$

Furthermore, the power spectral density function of a discrete-time stationary process is not exactly the same as that of its corresponding continuous-time stationary process. However, they are directly related by the following equation:

$$S_X(\omega) = \sum_{k=-\infty}^{\infty} \tilde{S}_X(\omega + 2\pi k), \qquad \omega \in [-\pi, \pi] \text{ and } k \in \mathbb{Z}. \tag{2.1.12}$$

2.1.4 Persistence

Persistence is a concept used to measure how close a single value of a time series is to the others. This concept is usually explained by either the autocovariance or the autocorrelation coefficient since both of them are used to measure the joint variability between values at two different points of time, which has been previously described. A time series of a discrete-time stationary process $\{X_t \mid t \in \mathbb{Z}\}$ is persistent if there exists strong positive autocorrelation coefficients at large lags τ. This can be explained by the summation of its autocovariance sequence or $s_{X,\tau}$ such that (McLeod and Hipel, 1978)

$$\sum_{\tau=-\infty}^{\infty} s_{X,\tau} = \infty, \qquad \tau \in \mathbb{Z}. \tag{2.1.13}$$

This definition also holds for the case of the autocovariance function of a continuous-time stationary process, by replacing the summation with an integral. This behaviour of the autocovariance is theoretically found in a long-memory process, and it is often

referred to as long-range dependence. On the other hand, if the summation in Equation (2.1.13) is finite, the process is weakly or short-range dependent. A special case of this behaviour is called anti-persistence such that the signs of the autocovariance (or the autocorrelation coefficients) alternate between positive and negative, and this results that the summation of autocovariance sequence over all lags τ is zero, i.e.,

$$\sum_{\tau=-\infty}^{\infty} s_{X,\tau} = 0, \qquad \tau \in \mathbb{Z}. \tag{2.1.14}$$

Building on these definitions, we will study various fractional and long-memory processes, and define their autocovariance functions (or sequences) and spectra as well as discuss whether such processes exhibit persistence or anti-persistence across different values of their parameters.

2.2 Fractional Brownian Motion

Fractional Brownian motion (FBM) was introduced by Mandelbrot and Van Ness (1968). They defined this fractional process in terms of a stochastic integral equation with the the Hurst exponent, denoted H, with a value between zero and one. This parameter was first used in hydrology for the time series representing the level of water flowing in a river, where its spectrum exhibits a power-law relationship (Hurst, 1951). It has then been widely used in various fields of study as a parameter measuring the degree of long-range dependence or long-memory behaviour of the time series or process with this power-law relationship. FBM, which we shall denote $B_H(t)$, is a nonstationary process due to its increasing variance over time. Several important characteristics of FBM include (1) its self-similar behaviour such that its autocovari-

ance function is identical to that of its time- and amplitude-rescaled version given by $\tilde{B}_H(t) \equiv \beta^H B_H(t/\beta)$, where $\beta \in \mathbb{R}^+$ is a scaling factor (Mandelbrot, 1985), (2) its degree of fractal dimension or small-scale roughness of the process (Mandelbrot, 1985; Falconer, 1990) which is regularly measured by the Hausdorff dimension $D = 2 - H$, ranging between one and two (Hausdorff, 1918; Adler, 1977; Gneiting and Schlather, 2004), and (3) its long-memory behaviour (Beran, 1994). FBM is always long-memory with any value of the Hurst exponent between zero and one. In addition, it is the only Gaussian stochastic process that is self-similar.

Definition 2.2.1 (General form of FBM)**:** FBM is a class of stochastic process with the following stochastic integral equation (Mandelbrot and Van Ness, 1968):

$$B_H(t) = \frac{A}{\Gamma(H+1/2)}\Bigg[\int_{-\infty}^{0}\left[(t-\tau)^{H-1/2} - (-\tau)^{H-1/2}\right]\mathrm{d}W(\tau) + \int_{0}^{t}(t-\tau)^{H-1/2}\mathrm{d}W(\tau)\Bigg], \qquad t \in \mathbb{R}, \tag{2.2.1}$$

where A is a parameter setting the level of process and $H \in (0,1)$. The initial condition on this stochastic integral equation is set as $B_H(0) = 0$. $W(\cdot)$ is called the Wiener process, and it is also known as ordinary Brownian motion. Thus, $\mathrm{d}W(\cdot)$ can be interpreted as continuous-time Gaussian white noise. When $H = 1/2$, Equation (2.2.1) becomes

$$B_{1/2}(t) = A\int_{0}^{t}\mathrm{d}W(\tau), \qquad t \in \mathbb{R}, \tag{2.2.2}$$

which recovers the Wiener process.

Definition 2.2.2 (Autocovariance function of FBM)**:** The nonstationary autocovariance function of FBM, which depends on both time t and lag τ, is defined as (Barton

and Poor, 1988; Lilly et al., 2017)

$$s_{B_H}(t,\tau) = \mathrm{Cov}[B_H(t), B_H(t+\tau)] = \frac{V_H}{2}A^2\big(|t+\tau|^{2H} + |t|^{2H} - |\tau|^{2H}\big), \tag{2.2.3}$$

where V_H is a normalising constant given by

$$V_H \equiv \frac{\Gamma(1-2H)\cos(\pi H)}{\pi H} = \frac{1}{\pi}\frac{\Gamma(H)\Gamma(1-H)}{\Gamma(2H+1)}. \tag{2.2.4}$$

With the substitution of $\tau = 0$ into Equation (2.2.3), the nonstationary variance of FBM takes the form

$$\mathrm{Var}[B_H(t)] = s_{B_H}(t,0) = V_H A^2 |t|^{2H}. \tag{2.2.5}$$

Hence, the variance increases without bound depending on time t, and at a rate determined by H.

Definition 2.2.3 (Power spectral density function of FBM)**:** The time-frequency Kirkwood-Rihaczek spectrum (Kirkwood, 1933; Rihaczek, 1968) of FBM, which allows for time-dependence in the power spectral density, is given by (Øigård et al., 2006)

$$S_{B_H}(t,\omega) = \pi V_H A^2 \,|t|^{2H}\,\delta(\omega) + \frac{A^2}{|\omega|^{2H+1}}\big(1 - e^{i\omega t}\big), \tag{2.2.6}$$

where $\delta(\cdot)$ is the Dirac delta function, and ω is the angular frequency in the units of radians per unit time. Lilly et al. (2017) presented its time-averaged spectrum by applying the moving average filter to the Kirkwood-Rihaczek spectrum with window size of T and then taking $T \to \infty$, i.e.,

$$\begin{aligned}\tilde{S}_{B_H}(\omega) &\equiv \lim_{T\to\infty}\frac{1}{T}\int_{t-T/2}^{t+T/2} S_{B_H}(u,\omega)\mathrm{d}u \\ &= \frac{A^2}{|\omega|^{2H+1}}.\end{aligned} \tag{2.2.7}$$

This spectrum is in the power-law form and clearly independent of time t. In addition, its rate of decay is controlled by a single parameter that is the Hurst exponent.

Although the general form of FBM is explained by the stochastic integral equation, we can simulate this process with the cumulative sum of its first-order difference process. Details of the simulation algorithm are provided later in Section 4.1. The plots of simulated FBMs with different true values of the Hurst exponent, defined as H_T, and each with the same length $n = 1000$ are given in Figure 2.2.1. In each plot, the time series is derived by sampling FBM at integer values of t only. Since a random sequence is used in the simulation for each plot (see Section 4.1), the estimated value of the Hurst exponent could be different from its true value, and this estimation challenge will be studied extensively in Chapters 3 and 4. From this figure, the more deterministic and straight-line behaviour is found in FBM with higher value of H_T as its time series is more persistent. Figure 2.2.2 shows the time-averaged spectra of FBMs from Equation (2.2.7) taking $T \to \infty$. The values of the Hurst exponent (H) are the same as H_T used in Figure 2.2.1, i.e., $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the parameter A is set as one. The spectra are all symmetric around the zero frequency due to the absolute value of ω in Equation (2.2.7), and this absolute value only exists in the denominator of this equation meaning that the spectra always approach infinity at the zero frequency. The steeper spectral line at all frequencies is observed from the process with higher value of H_T, as expected from this equation.

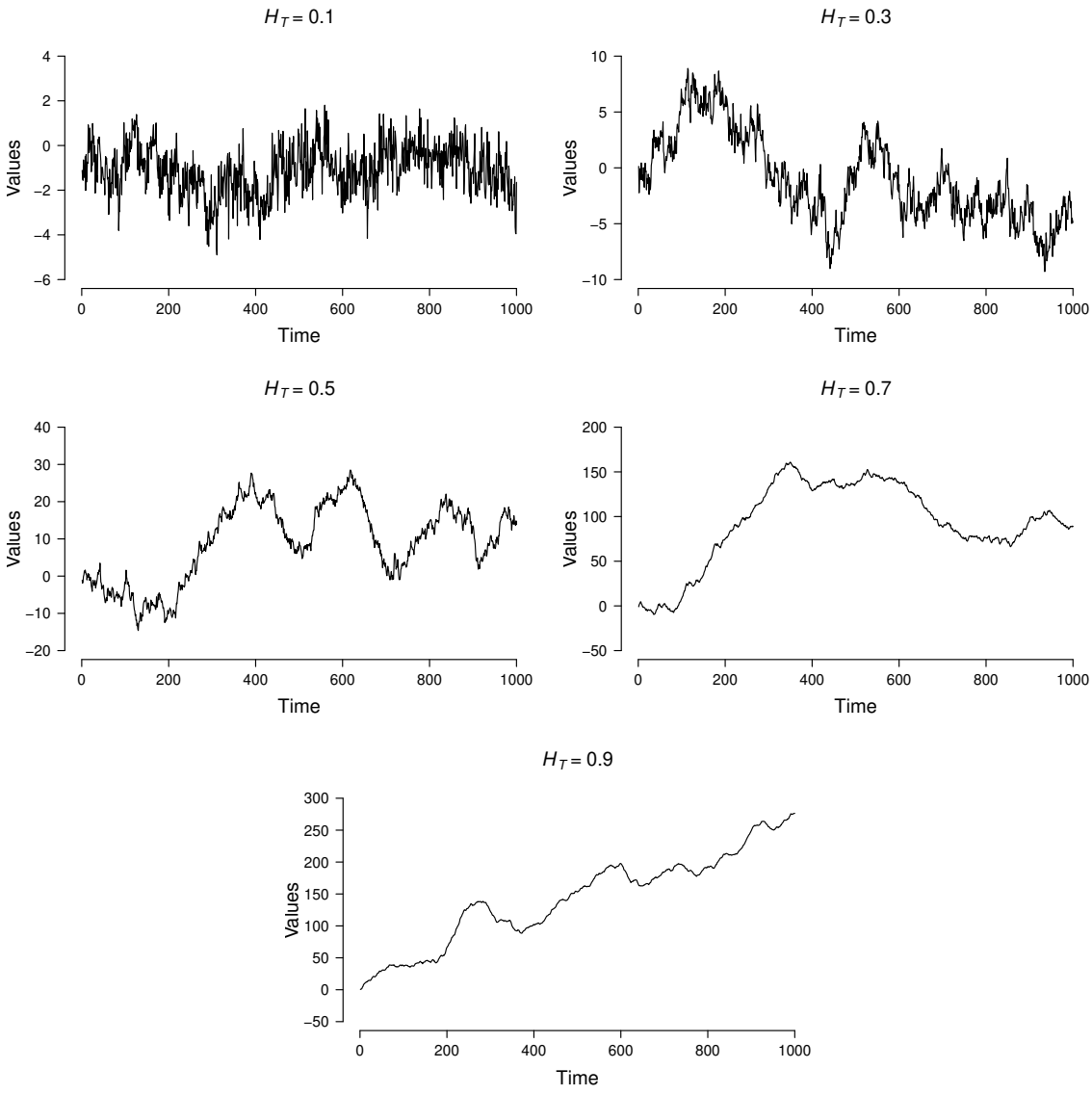

Figure 2.2.1: Simulated FBMs with different true values of the Hurst exponent such that $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Values are plotted at integer values of t only. Each time series is generated from the cumulative sum of simulated FGN (shown in Figure 2.3.1) with the same value of H_T. The true sample variance of FGN, denoted as σ_T^2, is set as one. The scale of the y-axis for each plot is different due to the nonstationary behaviour of each FBM.

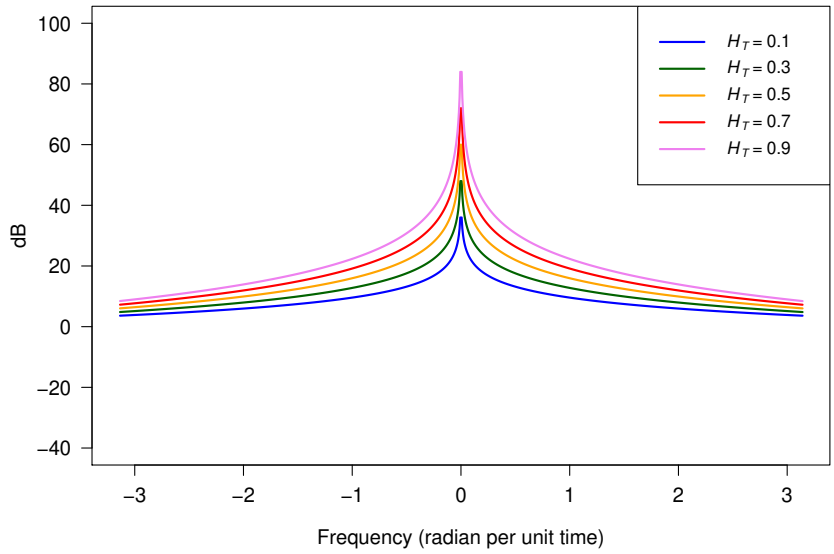

Figure 2.2.2: Time-averaged spectra of FBMs with $H_T = 0.1$ (blue), $H_T = 0.3$ (green), $H_T = 0.5$ (orange), $H_T = 0.7$ (red), and $H_T = 0.9$ (violet). All spectra are drawn from $A = 1$. The spectral values are in the units of decibels (dB).

2.3 Fractional Gaussian Noise

The first-order increment or difference process of regularly sampled observations of FBM is known as fractional Gaussian noise (FGN) (Mandelbrot and Van Ness, 1968). Unlike FBM, its variance and autocovariance function are both independent of time, making the process stationary. FGN is also controlled by the Hurst exponent ranging between zero and one. This process with $1/2 < H < 1$ has long-memory behaviour such that it is persistent with strong positive autocovariance. This autocovariance has a slow decay and it sums to infinity, in agreement with the definition of persistence given in Equation (2.1.13). Equivalently, its spectrum has power-law decay with the existence of a pole at the zero frequency, and this relationship with long memory will

be discussed later in Section 2.6. Gaussian white noise, corresponding to FGN with $H = 1/2$, on the other hand, is uncorrelated in time and has short memory with the same spectral value at all frequencies as this process always has a flat spectrum. The FGN with $0 < H < 1/2$ is an anti-persistent process such that the signs of the autocovariance alternate between positive and negative, and its spectrum approaches zero at the zero frequency (Robinson, 2003).

Definition 2.3.1 (General form of FGN)**:** The discrete-time FGN is defined as (Mandelbrot and Van Ness, 1968; Percival and Walden, 2000)

$$X_t \equiv B_H(t + \Delta) - B_H(t), \tag{2.3.1}$$

where $B_H(t)$ was defined in Equation (2.2.1), and $\Delta \in \mathbb{R}^+$ is the sampling interval.

Definition 2.3.2 (Autocovariance sequence of FGN)**:** The autocovariance sequence of FGN is stationary or independent from time t, and this is given by

$$s_{X,\tau} = \text{Cov}[X_t, X_{t+\tau}] = \frac{V_H}{2} A^2 \big(|\tau + \Delta|^{2H} - 2|\tau|^{2H} + |\tau - \Delta|^{2H}\big). \tag{2.3.2}$$

Henceforth, the value of Δ is set to one for simplicity, and without loss of generality. The variance of $\{X_t\}$ is derived by setting $\tau = 0$ into Equation (2.3.2), i.e.,

$$\text{Var}[X_t] \equiv \sigma_X^2 = V_H A^2. \tag{2.3.3}$$

Because A is the parameter setting the level of FBM, we consider σ_X^2 as a parameter setting the variance and the autocovariance sequence of FGN. Thus, the autocovariance sequence of FGN can be specified by two parameters: H and σ_X^2, where A is then found using the relationship $A = \sigma_X / \sqrt{V_H}$.

According to Equation 2.3.1, $\{X_t\}$ with a finite length of $t = 1, 2, \ldots, n$ can be expressed as

$$X_t \equiv \sum_{l=0}^{1} a_l B_{t-l}, \qquad l \in \mathbb{Z}, \tag{2.3.4}$$

where $B_{t-l} = B_H(t-l)$. As we have set $\Delta = 1$ for simplicity, the impulse response sequence of a linear filter $\{a_l\}$ is simply given by $a_0 = -1$, $a_{-1} = 1$, and $a_k = 0$ for any integer $k \neq 0, -1$. A transfer function of a linear filter of width W is defined as

$$A(\omega) \equiv \sum_{l=0}^{W-1} a_l e^{-i\omega l}. \tag{2.3.5}$$

Thus, the transfer function of our impulse response sequence can be written as $A(\omega) = (\cos\omega - 1) - i\sin\omega$. The power spectral density function of $\{X_t\}$ is equivalent to the multiplication of the spectrum of the regularly-sampled FBM with $\Delta = 1$ and the squared gain function of this linear filter given by $\mathcal{D}(\omega) = |A(\omega)|^2 = 4\sin^2(\omega/2)$ (Percival and Walden, 2000). This allows us to express the spectrum of FGN from the time-averaged spectrum of FBM in Equation (2.2.7).

Definition 2.3.3 (Power spectral density function of FGN)**:** The spectrum of FGN is given by

$$S_X(\omega) = 4\sin^2\left(\frac{\omega}{2}\right) \sum_{j=-\infty}^{\infty} \frac{A^2}{|\omega + 2\pi j|^{2H+1}}, \qquad -\pi \leq \omega \leq \pi. \tag{2.3.6}$$

This spectrum is defined only within the range of discrete frequencies less than or equal to the Nyquist frequency (or half of the sampling rate $= 1/2 \times 2\pi = \pi$ radians per unit time) in magnitude. The infinite summation occurs because of the phenomenon of *aliasing*. Aliasing occurs when a discrete-time process (such as FGN) is regularly sampled from a continuous-time process (such as FBM) and the spectrum is now

only defined at or below the Nyquist frequency rather than across all real-valued frequencies. When this sampling occurs, the power at frequencies above the Nyquist *aliases* to frequencies below the Nyquist (separated by modulus 2π) and is observed together in aggregation. The infinite summation in Equation (2.3.6) has no exact analytical or closed-form solution, and one option is to ignore the aliased frequencies and simply just take the value of $j = 0$ in Equation (2.3.6)—we will call this the *unaliased* FGN spectrum. A better approximation, however, can be obtained using the Euler-Maclaurin formula such that (Percival and Walden, 2000)

$$S_X(\omega) \approx 4\sin^2\left(\frac{\omega}{2}\right)\left(\frac{A^2}{(2\pi)^{2H+1}}\right)\left(\sum_{j=-M}^{M}\frac{1}{|f+j|^{2H+1}} + \sum_{l=-1,1}\left[\frac{1}{2H(lf+M+1)^{2H}} - \frac{(2H+1)(2H+2)(2H+3)}{720(lf+M+1)^{2H+4}} + \frac{2H+1}{12(lf+M+1)^{2H+2}} + \frac{1}{2(lf+M+1)^{2H+1}}\right]\right), \qquad -\pi \le \omega \le \pi, \quad (2.3.7)$$

where $f = \omega/2\pi$ and M is an arbitrary positive integer. Percival and Walden (2000) suggested $M = 100$ yielding good approximation in practice. We will call the approximated spectrum of Equation (2.3.7) as the *aliased* FGN spectrum, and we will compare with different values of $M = 0$ and $M = 100$ later in Chapters 3 and 4 for the estimation of H and σ_X^2.

Figure 2.3.1 shows several examples of simulated FGNs with different true values of the Hurst exponent (the same set of H_T used in the simulation of FBMs), all with the same true value of the sample variance, denoted as $\sigma_T^2 = 1$. From this figure, we observe that FGN is more persistent with higher H_T, while there is a negative autocorrelation of observed values of time series, or anti-persistence, in FGN

with lower H_T. These result in different characteristics of their spectra as shown in Figure 2.3.2. Because there is no closed-from expression for the spectrum of a discrete-time FGN, we used its approximated aliased form derived from the Euler-Maclaurin formula in Equation (2.3.7) with $M = 100$. The spectrum of each persistent and long-memory process with $H_T > 0.5$ approaches infinity (∞ dB) at the zero frequency, whereas the spectrum of each anti-persistent process with $H_T < 0.5$ approaches zero ($-\infty$ dB) at this frequency. The spectrum of Gaussian white noise with $H_T = 0.5$ is constant and its amplitude is $10 \log_{10}(1/2\pi) = -7.9818$ dB.

2.4 The Fractionally Differenced Process

The discrete-time fractionally differenced (FD) process has been often studied alongside FGN and FBM. Instead of the Hurst exponent, this process has the parameter $d \in \mathbb{R}$ to measure the degree of its long-range dependence. This parameter also represents the order of fractional differencing required for the process to become Gaussian white noise, which is a collection of independent and identically distributed (i.i.d.) random variables with zero mean and finite variance.

Definition 2.4.1 (General form of the FD process)**:** Let $\{\varepsilon_t\}$ be a white noise process with a variance of σ_ε^2. This process takes the form

$$\varepsilon_t = (1 - B)^d Y_t, \tag{2.4.1}$$

where $\{Y_t\}$ is the FD process, and B is the backshift operator with $B^m Y_t = Y_{t-m}$.

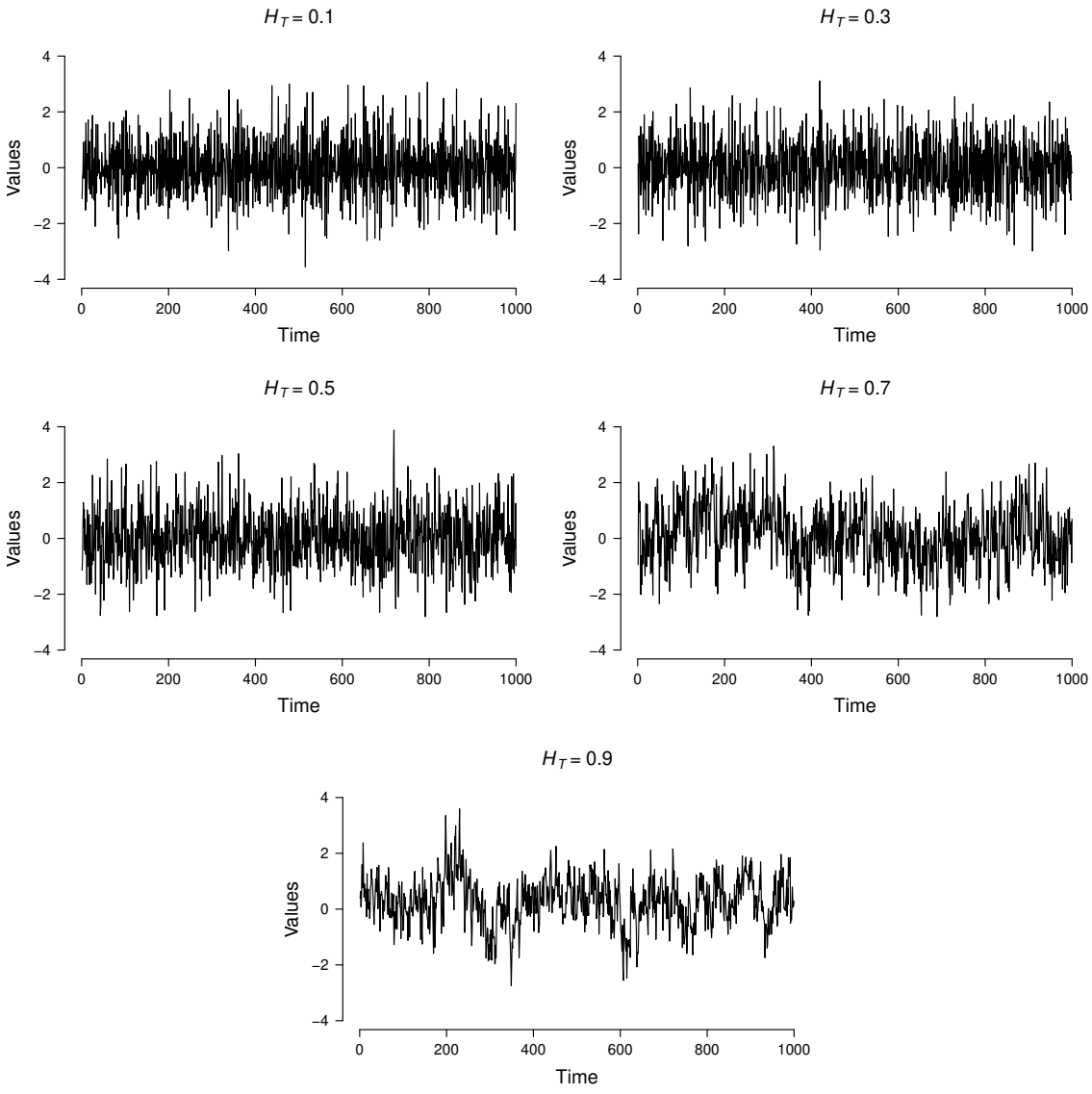

Figure 2.3.1: Simulated FGNs with the same true value of the sample variance, $\sigma_T^2 = 1$, but different true values of the Hurst exponent such that $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Values are plotted at integer values of t only. Details of the simulation algorithm are provided later in Section 4.1.

For any real value of d, the term $(1 - B)^d$ can be expanded using the Binomial series:

$$\begin{aligned}(1-B)^d &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \\ &= 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \dots \end{aligned} \tag{2.4.2}$$

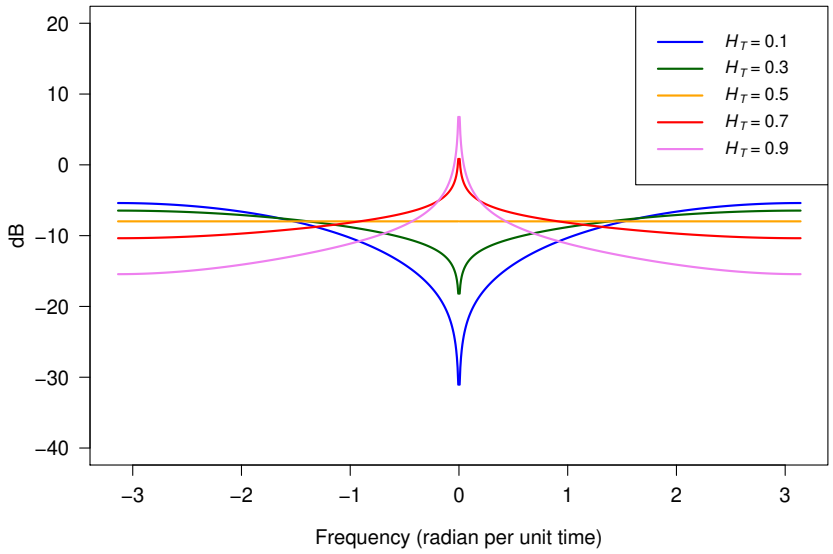

Figure 2.3.2: Spectra of FGNs with $H_T = 0.1$ (blue), $H_T = 0.3$ (green), $H_T = 0.5$ (orange), $H_T = 0.7$ (red), and $H_T = 0.9$ (violet) by using the approximated form in Equation (2.3.7) with $M = 100$. All spectra are drawn from $\sigma_T^2 = 1$.

The general form of $\{Y_t\}$ is therefore given by

$$\begin{aligned} Y_t &= (1-B)^{-d}\varepsilon_t \\ &= \left(1 + dB + \frac{d(d+1)}{2!}B^2 + \frac{d(d+1)(d+2)}{3!}B^3 + \dots\right)\varepsilon_t, \end{aligned} \tag{2.4.3}$$

which represents an infinite moving average or MA(∞) process. This process is controlled by two parameters: d and σ_ε^2.

Definition 2.4.2 (Autocovariance sequence of the FD process)**:** The FD process is only stationary when $d < 1/2$. For $d = 0$, $\{Y_t\}$ is exactly the same as the white noise process with the autocovariance of σ_ε^2 at lag zero and zero elsewhere. For other values of $d < 1/2$, the autocovariance sequence is given by (Percival and Walden, 2000)

$$s_{Y,\tau} = \mathrm{Cov}[Y_t, Y_{t+\tau}] = \frac{\sigma_\varepsilon^2 \sin(\pi d)\Gamma(1-2d)\Gamma(\tau+d)}{\pi\Gamma(\tau+1-d)}, \tag{2.4.4}$$

where $\Gamma(\cdot)$ is the gamma function. The variance of $\{Y_t\}$ is derived by setting $\tau = 0$, and hence it takes the form

$$\mathrm{Var}[Y_t] = \frac{\sigma_\varepsilon^2 \sin(\pi d)\Gamma(1-2d)\Gamma(d)}{\pi\Gamma(1-d)} = \frac{\sigma_\varepsilon^2\Gamma(1-2d)}{\Gamma^2(1-d)}. \tag{2.4.5}$$

Definition 2.4.3 (Power spectral density function of the FD process)**:** The squared gain function for the d^{th} order difference filter is given by $\mathcal{D}^d(\omega) = (2\sin(\omega/2))^{2d}$, and thus the power spectral density function of $\{Y_t\}$ is

$$S_Y(\omega) = \frac{\sigma_\varepsilon^2}{(2\sin(\omega/2))^{2d}}, \qquad -\pi \le \omega \le \pi. \tag{2.4.6}$$

This can be rewritten as

$$\begin{aligned} S_Y(\omega) &= \frac{\sigma_\varepsilon^2}{(2\sin(\omega/2))^{2d}} \\ &= \frac{\sigma_\varepsilon^2}{(2(1-\cos(\omega))^d} \\ &= \frac{\sigma_\varepsilon^2}{((1-\cos(\omega))^2 + \sin^2(\omega))^d} \\ &= \frac{\sigma_\varepsilon^2}{|1-e^{i\omega}|^{2d}}, \qquad -\pi \le \omega \le \pi, \end{aligned} \tag{2.4.7}$$

which is the same as that given by Adenstedt (1974). This spectrum is an even function for all real values of d. It is the Fourier transform of the autocovariance sequence in Equation (2.4.4) for $d < 1/2$. For $d \ge 1/2$, the spectrum is non-integrable as the variance and the autocovariance are nonstationary.

2.5 Pure Power Law Process

Previously, the power spectral density functions of FBM, FGN, and the FD process have been given. One common feature of all these functions is that they are considered to have a power-law function *at low frequencies* (see Equations (2.2.7), (2.3.6),

and (2.4.6) with the approximation of $\sin(\omega/2) \approx \omega/2$ when $\omega \to 0$). If this power-law relationship were to be extended to all observed frequencies, we would call the process with such behaviour as the pure power law (PPL) process which is a discrete-time process as we shall now define.

Definition 2.5.1 (Power spectral density function of the PPL process)**:** Let $\{P_t\}$ be the PPL process. Its power spectral density function is given by

$$S_P(\omega) = C|\omega|^\gamma, \qquad -\pi \leq \omega \leq \pi, \tag{2.5.1}$$

where $C > 0$ and $\gamma \in \mathbb{R}$. $\{P_t\}$ is either stationary with $\gamma > -1$ or nonstationary with $\gamma \leq -1$. This spectrum can be compared with those of other discrete-time fractional processes. The spectrum of FGN at low frequencies is close to the spectrum of the PPL process with $\gamma = 1 - 2H$. This relationship is derived from Equation (2.3.6) with the approximation of $\sin(\omega/2) \approx \omega/2$ at low frequencies and by neglecting the aliasing terms. Such an approximation can also be applied to the spectrum of FD process at low frequencies from Equation (2.4.6), and this results in its approximated form of $C|\omega|^\gamma$, where $C = \sigma_\varepsilon^2$ and $\gamma = -2d$. Unlike other fractional processes, there is no exact analytical or close-form solution of the autocovariance sequence of the PPL process.

Definition 2.5.2 (Autocovariance sequence of the PPL process)**:** There is no closed-form expression of the autocovariance sequence of the PPL process. However, the autocovariance is equivalent to the inverse Fourier transform of the power spectral density function, i.e.,

$$s_{X,\tau} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(\omega)e^{i\omega\tau}\mathrm{d}\omega = \frac{C}{2\pi}\int_{-\pi}^{\pi} |\omega|^\gamma e^{i\omega\tau}\mathrm{d}\omega. \tag{2.5.2}$$

The approximation of this integral can be performed using techniques such as Riemann approximation with carefully selected bin widths, as developed in Grainger et al. (2021), which we refer to the interested reader towards for more discussion.

2.6 Long-Memory Parameters

In this section, we provide conditions on parameters for each process discussed in this chapter to be long-memory or long-range dependent. These depend on either its autocovariance function (or sequence) or power spectral density function, which is given as follows:

Let $\{Q_t\}$ be a discrete-time stationary fractional process. The process is long-memory if it satisfies (Beran, 1994)

$$\lim_{\tau \to \infty} \frac{s_{Q,\tau}}{c_1 \tau^{\beta}} = 1, \qquad -1 < \beta < 0, \qquad c_1 > 0, \tag{2.6.1}$$

or

$$\lim_{\omega \to 0} \frac{S_Q(\omega)}{c_2 |\omega|^{-\beta-1}} = 1, \qquad -1 < \beta < 0, \qquad c_2 > 0, \tag{2.6.2}$$

where $s_{Q,\tau}$ and $S_Q(\omega)$ are the autocovariance sequence and the spectrum of $\{Q_t\}$, respectively. Therefore, at increasing lags, the autocovariance sequence of a long-memory process has slow power-law decay such that it is not absolutely integrable. Furthermore, the spectrum of a long-memory process exhibits power-law decay at low frequencies. This results in a pole at the zero frequency, and such behaviour is in contrast to the spectrum of a fractional process with anti-persistence having its spectrum approach zero at the zero frequency.

We now formally establish the long-memory properties of each of the discrete-time stationary fractional processes.

2.6.1 Fractional Gaussian Noise

First, we assume that $\{Q_t\}$ is a stationary FGN, whose properties were given in Section 2.3. Its autocovariance sequence with $\Delta = 1$ in Equation (2.3.2) is rewritten, by taking $\tau \to \infty$ and making use of the binomial series expansion, as

$$\begin{aligned} s_{Q,\tau} &= \frac{V_H}{2} A^2 |\tau|^{2H} \left[\left| 1 + \frac{1}{\tau} \right|^{2H} - 2 + \left| 1 - \frac{1}{\tau} \right|^{2H} \right] \\ &\approx \frac{V_H}{2} A^2 |\tau|^{2H} \left[1 + \frac{2H}{\tau} + \frac{H(2H-1)}{\tau^2} - 2 + 1 - \frac{2H}{\tau} + \frac{H(2H-1)}{\tau^2} \right] \\ &= \frac{V_H}{2} A^2 |\tau|^{2H} \left[\frac{2H(2H-1)}{\tau^2} \right] \\ &= V_H A^2 H(2H-1) \tau^{2H-2}, \qquad \tau \to \infty. \end{aligned} \tag{2.6.3}$$

With the substitution of this autocovariance sequence into Equation (2.6.1), the process $\{Q_t\}$ is long-memory if and only if $-1 < 2H - 2 < 0$, i.e., $1/2 < H < 1$, and its autocovariance sequence decays like τ^{2H-2} as $\tau \to \infty$. The power spectral density function of FGN in Equation (2.3.6) at low frequencies can be approximated to the power-law form of $C|\omega|^{1-2H}$, where C is a positive constant. According to Equation (2.6.2), this spectrum belongs to a long-memory process if and only if $-1 < (1 - 2H) < 0$ or $1/2 < H < 1$, and this is the same condition as derived from the autocovariance sequence.

2.6.2 The Fractionally Differenced Process

Next, $\{Q_t\}$ is assumed to be a stationary FD process with $d < 1/2$, which was previously discussed in Section 2.4. Its autocovariance sequence in Equation (2.4.4) at large lag τ can be approximated by Stirling's formula and this is given by

$$s_{Q,\tau} \approx \frac{\sigma_\varepsilon^2 \sin(\pi d)\Gamma(1-2d)}{\pi}\tau^{2d-1}, \qquad \tau \to \infty. \tag{2.6.4}$$

The process $\{Q_t\}$ is long-memory if and only if $-1 < 2d-1 < 0$, i.e., $0 < d < 1/2$, and its autocovariance sequence decays like τ^{2d-1} as $\tau \to \infty$. The power spectral density function of the FD process in Equation (2.4.6) takes the approximated form of $C|\omega|^{-2d}$ with a positive constant C as $\omega \to 0$, and this also provides a result that the process is long-memory if and only if $0 < d < 1/2$.

Then we consider a nonstationary FD process $\{Q'_t\}$ with $d' \geq 1/2$ (d' is its order of differencing). The stationary process $\{Q_t\}$ is the k^{th} order difference of $\{Q'_t\}$ with $k > 0$, and hence its spectrum is given by

$$S_Q(\omega) \equiv S_{Q'}(\omega) \cdot \mathcal{D}^k(\omega) = \frac{\sigma_\varepsilon^2}{(2\sin(\omega/2))^{2(d'-k)}}, \qquad -\pi \leq \omega \leq \pi, \tag{2.6.5}$$

where $\mathcal{D}^k(\omega)$ is the squared gain function of the k^{th} order difference filter, i.e., $\mathcal{D}^k(\omega) = (2\sin(\omega/2))^{2k}$. Hence, the spectrum of $\{Q'_t\}$ is in the form of

$$S_{Q'}(\omega) = \frac{\sigma_\varepsilon^2}{(2\sin(\omega/2))^{2d'}}, \qquad -\pi \leq \omega \leq \pi. \tag{2.6.6}$$

Percival and Walden (2000) defined the parameter d' as the long-memory parameter for nonstationary FD process. This process is always long-memory with $d' \geq 1/2$. The power spectral density function of stationary and nonstationary FD processes (Equations (2.4.6) and (2.6.6)) can be used to deduce that the spectrum of the discrete-time

long-memory process at low frequencies is in the form of $C|\omega|^{\gamma}$, where $C > 0$ and $\gamma < 0$. This is equivalent to the spectrum of the PPL process shown in Section 2.5.

2.6.3 Fractional Brownian Motion

The power spectral density function of $\{B_t\}$, which we define to be a regularly-sampled (or discretely-sampled) FBM with a sampling rate of $\Delta = 1$, is similar to that of FGN in Equation (2.3.6), but without the squared gain function for the difference filter. Because the spectral values at low frequencies are much greater than high frequencies especially those above the Nyquist frequency, we can approximate the spectrum by neglecting the folding from such high frequencies, i.e., $S_B(\omega) \approx A^2/|\omega|^{2H+1}$. Hence, the process $\{B_t\}$ is long-memory if and only if $-1 - 2H < 0$. As $0 < H < 1$, this concludes that FBM with discrete sampling is always long-memory for any value of H.

2.6.4 Comparison

All specific conditions on parameters for the fractional processes discussed in this chapter to be long-memory, stationary, and nonstationary, are summarised in Table 2.6.1. The relationship of these parameters for long-memory processes can be given by either $H = d + 1/2 = (1 - \gamma)/2$ (stationary) or $H = d - 1/2 = (-1 - \gamma)/2$ (nonstationary) depending on whether or not the processes are stationary. The PPL process with $\gamma = -1$ is sometimes referred to as $1/f$ or pink noise. Its spectrum is in the form of $C/|\omega|$, and this behaviour is also found in the spectrum of the FD process

with $d = 1/2$ at low frequencies. However, this form is unavailable for FGN and FBM as the Hurst exponent cannot be exactly one or zero, but the limiting case between FGN with H close to one and FBM with H close to zero can be considered to behave like $1/f$ or pink noise.

Table 2.6.1: The range of parameters satisfying the conditions of long memory, stationarity, and nonstationarity, for four different fractional processes.

Process	Long Memory	Stationarity	Nonstationarity
FBM (Discrete sampling)	$0 < H < 1$	$-$	$0 < H < 1$
FGN	$1/2 < H < 1$	$0 < H < 1$	$-$
FD	$d > 0$	$d < 1/2$	$d \geq 1/2$
PPL	$\gamma < 0$	$\gamma > -1$	$\gamma \leq -1$

Figure 2.6.1 illustrates examples of power spectral density functions of stationary and nonstationary long-memory processes with $H = 0.7$. For each case (stationary or nonstationary), their spectral lines are almost the same at low frequencies. In addition, they approach infinity when the frequency is close to zero. The nonstationary processes have much higher spectra at low frequencies than the stationary processes since their observed values are closer to each other or have higher degree of positive autocorrelation, yielding smoother time series that are characteristic of low-frequency dominated time series.

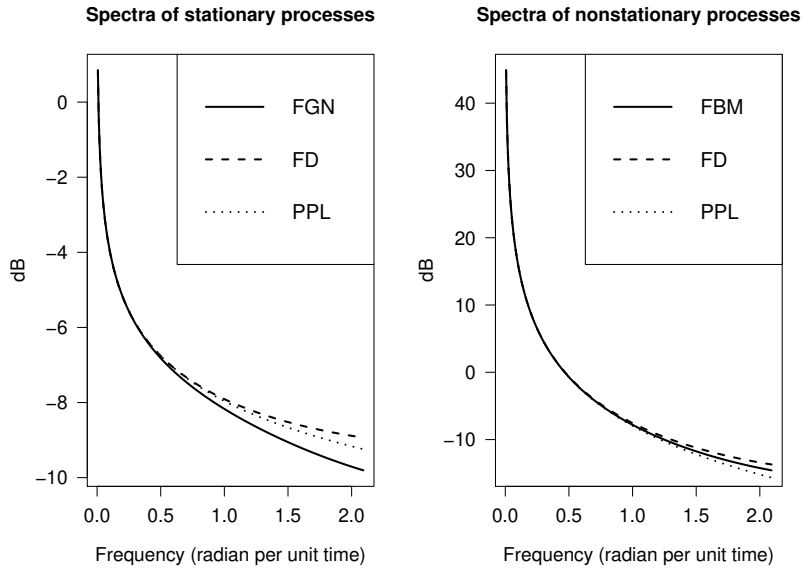

Figure 2.6.1: Power spectral density functions of stationary (left) and nonstationary (right) long-memory processes. The values of fractional parameters are $\{H, d, \gamma\} = \{0.7, 0.2, -0.4\}$ for stationary processes, and $\{H, d, \gamma\} = \{0.7, 1.2, -2.4\}$ for nonstationary processes.

Chapter 3

Estimation Methods

In this chapter, we propose several parametric methods for the estimation of long-memory parameters. Section 3.1 provides the details of the maximum likelihood estimator (MLE). Despite its consistency and asymptotic efficiency, this estimator is not implemented in practice in this thesis due to its high computational cost. Instead, its approximated version in the frequency domain called the Whittle likelihood estimator (WLE) will be used, and this is given in Section 3.2. The debiased Whittle likelihood estimator (DWLE), which was recently proposed by Sykulski et al. (2019), will also be used to reduce the bias of estimates from the WLE. Although Sykulski et al. (2019) used the DWLE with some short-memory processes, we will instead investigate using it with long-memory processes in this thesis. This debiased estimator is discussed in Section 3.3. Next, the periodogram is the asymptotically unbiased estimator of the spectrum, and a slope of its linear regression line at low frequencies in log-log space can be used to estimate a long-memory parameter due to the power-law relationship. This method is called the log-periodogram regression estimator (LPRE), and it is de-

tailed in Section 3.4. All these four sections are explained in the context of estimating H and σ_X^2 for stationary FGN. However, the principles and techniques apply to other processes with some basic modifications. In particular, the details of the estimation of parameters with the stationary FD process and a discretely sampled FBM process are given in Sections 3.5 and 3.6, respectively. Because the periodogram is a biased spectral estimator of the spectrum, especially with nonstationary processes such as FBM, data tapering is proposed to reduce such bias, and this technique is described in Section 3.7.

3.1 Maximum Likelihood Estimator

Let $X = \{X_t\}_{t=1}^n$ be discrete-time FGN with zero mean and unit variance. The likelihood function or joint probability density function is derived from the multivariate normal distribution of all random variables in X, i.e.,

$$L\left(H, \sigma_X^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} X^T \mathbf{\Sigma}^{-1} X \right), \tag{3.1.1}$$

where $\mathbf{\Sigma}$ is an $n \times n$ autocovariance matrix with entries $[\mathbf{\Sigma}]_{ij} = s_{X,|i-j|}$, which can be expressed as $s_{X,|i-j|} = \sigma_X^2(|i-j+1|^{2H} - 2|i-j|^{2H} + |i-j-1|^{2H})/2$, thus consisting of both H and σ_X^2. The superscripts "T" and "-1" denote the transpose and the matrix inverse, respectively, while $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$. The log-likelihood function of Equation (3.1.1) is given by

$$\ell\left(H, \sigma_X^2\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\big(\log|\mathbf{\Sigma}| + X^T \mathbf{\Sigma}^{-1} X\big). \tag{3.1.2}$$

Then the maximum likelihood estimation of parameters in a set of $\boldsymbol{\theta} = \{\theta_1, \theta_2\} = \{H, \sigma_X^2 \,|\, 0 < H < 1, \sigma_X^2 > 0\}$ is

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}\right) = \arg\min_{\boldsymbol{\theta}} \big(\log|\boldsymbol{\Sigma}| + X^T \boldsymbol{\Sigma}^{-1} X \big), \tag{3.1.3}$$

where $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is a set of estimates. The MLE is consistent such that the estimates converge in probability to their true values, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} \xrightarrow{p} \boldsymbol{\theta}_{\text{T}}, \tag{3.1.4}$$

where $\boldsymbol{\theta}_{\text{T}} = \{H_T, \sigma_T^2\}$ is the set of true values of parameters. In addition, the maximum likelihood estimates are $\sqrt{n}$-consistent with the convergence in distribution and asymptotically efficient, i.e.,

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{\text{T}}\big) \xrightarrow{d} \mathcal{N}\big(0, \boldsymbol{\mathcal{F}}^{-1}\big), \tag{3.1.5}$$

where $\boldsymbol{\mathcal{F}}$ is the Fisher information matrix. This matrix is defined as the negative of the expected value of the Hessian matrix $\boldsymbol{\mathcal{H}}$ with its entry being the second partial derivatives of the log-likelihood function given the true parameters, i.e.,

$$[\boldsymbol{\mathcal{F}}]_{ij} = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}\right]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \theta_j} \ell\left(\boldsymbol{\theta}_{\text{T}}\right)\right], \tag{3.1.6}$$

where $1 \leq i \leq 2$ and $1 \leq j \leq 2$.

3.2 Whittle Likelihood Estimator

Although the MLE is consistent and asymptotically efficient, a huge disadvantage of this estimator is its very high computational cost of estimation especially from

the matrix inversion in the log-likelihood function of Equation (3.1.2), which requires at least $\mathcal{O}(n^2)$ operations in general. To reduce this cost, its approximated version called the Whittle likelihood estimator (WLE) was proposed by Whittle (1953). The log-likelihood function of this estimator in the frequency domain is given by

$$\ell_{\text{WLE}}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}}\left(\log\left(S_X(\omega|\boldsymbol{\theta})\right) + \frac{I(\omega)}{S_X(\omega|\boldsymbol{\theta})}\right), \tag{3.2.1}$$

where $\boldsymbol{\Omega}$ is a finite set of discrete Fourier frequencies, i.e., $\boldsymbol{\Omega} \equiv \{\omega_j\}_{j=1}^{n}$ for each $\omega_j = 2\pi(j - \lceil n/2\rceil)/n$. Thus, the range of these frequencies is between $-\pi$ and π radians per unit time. $S_X(\omega|\boldsymbol{\theta})$ is the power spectral density function of X given the set of $\boldsymbol{\theta} = \{H, \sigma_X^2\}$, and $I(\omega)$ is the asymptotically unbiased estimator of the power spectral density function called the periodogram which is defined as (Percival and Walden, 2000)

$$I(\omega) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t \exp(-i\omega t)\right|^2, \qquad \omega \in \boldsymbol{\Omega}. \tag{3.2.2}$$

The use of the periodogram in the likelihood of Equation (3.2.1) removes the need to invert a covariance matrix as performed with the maximum likelihood in Equation (3.1.1). The periodogram of Equation (3.2.2) can be computed in $\mathcal{O}(n\log n)$ operations by making use of the Fast Fourier transform (FFT), thus making the Whittle likelihood significantly faster to compute than the maximum likelihood, which is a key reason for its wide appeal in many time series applications. The Whittle likelihood estimates of $\boldsymbol{\theta}$ is a set of values of parameters maximising the log-likelihood function in Equation (3.2.1). This is equivalent to finding

$$\hat{\boldsymbol{\theta}}_{\text{WLE}} = \arg\min_{\boldsymbol{\theta}} \sum_{\omega\in\boldsymbol{\Omega}}\left(\log\left(S_X(\omega|\boldsymbol{\theta})\right) + \frac{I(\omega)}{S_X(\omega|\boldsymbol{\theta})}\right), \tag{3.2.3}$$

where $\hat{\boldsymbol{\theta}}_{\text{WLE}}$ is a set of estimates from the WLE. The WLE has also been shown to be consistent and asymptotically efficient for a Gaussian process with long-memory time series (Robinson, 2003), and a range of non-Gaussian processes, widening its appeal in practice (Sykulski et al., 2019).

For a long-memory FGN with $1/2 < H_T < 1$, its power spectral density function approaches infinity at the zero frequency. Therefore, the summation in Equation (3.2.1) must exclude the zero frequency from $\boldsymbol{\Omega}$, which is a small disadvantage of the WLE as the information contained in the periodogram at zero frequency is discarded, slightly adding to the variance of the parameter estimates versus maximum likelihood. Furthermore, because there is no closed-form expression of the spectrum of FGN, an approximation must be used (see Equations (2.3.6) and (2.3.7)). We propose the use of three different functions which we now detail. The first two functions are approximated aliased spectral density functions in Equation (2.3.7) with $M = 100$ and with $M = 0$, respectively. We use different values of M to observe the trade-off between bias (lower bias when M is high) and computational time (lower computational time when M is low). The other function is obtained from the removal of the infinite sum of the exact spectral density function in Equation (2.3.6) by setting $j = 0$, and this results in using the unaliased spectral density function. The motivation for considering all three forms is that they will have different computational complexities to compute (with the unaliased spectrum being the fastest), and therefore helps to understand how much correction to aliasing is necessary and reasonable to compute. The estimation results from all these different forms of approximated spectrum will be provided in Chapter 4.

3.3 Debiased Whittle Likelihood Estimator

The debiased Whittle likelihood estimator (DWLE) was proposed by Sykulski et al. (2019) to reduce the bias of estimates from the WLE. This estimator adjusts the log-likelihood function of the WLE by replacing its power spectral density function with the expected periodogram, and this is given by

$$\ell_{\mathrm{DWLE}}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}}\left(\log\left(\bar{S}(\omega|\boldsymbol{\theta})\right) + \frac{I(\omega)}{\bar{S}(\omega|\boldsymbol{\theta})}\right), \tag{3.3.1}$$

where $\bar{S}(\omega|\boldsymbol{\theta})$ is the expected periodogram given the set of parameters. The expected periodogram is defined as

$$\bar{S}(\omega|\boldsymbol{\theta}) \equiv \mathbb{E}\left[I(\omega)\right] = \int_{-\pi}^{\pi} S_X(\nu|\boldsymbol{\theta})\mathcal{F}(\omega-\nu)\,d\nu, \tag{3.3.2}$$

which is the convolution of the power spectral density function and the Fejér kernel. This kernel is given by

$$\mathcal{F}(\omega) \equiv \frac{\sin^2(n\omega/2)}{n\sin^2(\omega/2)}. \tag{3.3.3}$$

Because the convolution for the expected periodogram has no closed-form expression, it is transformed to the point-wise multiplication in the time domain as (Percival and Walden, 1993)

$$\begin{aligned}
\bar{S}(\omega|\boldsymbol{\theta}) &= \sum_{\tau=-(n-1)}^{n-1} \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n-|\tau|} X_t X_{t+|\tau|}\right] e^{-i\omega\tau} \\
&= \sum_{\tau=-(n-1)}^{n-1}\left(1-\frac{|\tau|}{n}\right)s_{X,\tau}e^{-i\omega\tau} \\
&= 2\cdot\mathrm{Re}\left(\sum_{\tau=0}^{n-1}\left(1-\frac{\tau}{n}\right)s_{X,\tau}e^{-i\omega\tau}\right) - \sigma_X^2,
\end{aligned} \tag{3.3.4}$$

where $\mathrm{Re}(\cdot)$ indicates the real part. Hence, the expected periodogram can be computed from the fast Fourier transform of the multiplication of the autocovariance

sequence and the triangular function. Equation (3.3.4) also shows that aliasing is accounted for by the discrete sampling of the autocovariance sequence $s_{X,\tau}$, and an effect called spectral leakage (which we shall discuss further in Section 3.7) is accounted for by the truncation of the infinite sum and the insertion of the triangular function $1 - \tau/n$, and both corrections are performed in a single operation by Equation (3.3.4). This ensures that the DWLE can also be computed in $\mathcal{O}(n \log n)$ operations (as is the case with the WLE). However, for the DWLE, two FFTs are required to computed Equation (3.3.1) — one to compute $I(\omega)$ and the other to compute the expected periodogram $\bar{S}(\omega|\boldsymbol{\theta})$. Overall, the method is still fast to compute in comparison to the exact maximum likelihood such that its $\mathcal{O}(n \log n)$ complexity significantly outperforms the $\mathcal{O}(n^2)$ complexity of the MLE (in terms of reduced CPU time), as demonstrated in Sykulski et al. (2019) and Grainger et al. (2021).

The debiased Whittle likelihood estimates maximise the log-likelihood function in Equation (3.3.1). This is the same as optimising the following equation:

$$\hat{\boldsymbol{\theta}}_{\text{DWLE}} = \arg\min_{\boldsymbol{\theta}} \sum_{\omega \in \boldsymbol{\Omega}} \left(\log\left(\bar{S}(\omega|\boldsymbol{\theta})\right) + \frac{I(\omega)}{\bar{S}(\omega|\boldsymbol{\theta})} \right). \tag{3.3.5}$$

The expected periodogram of FGN is finite at all discrete Fourier frequencies whether or not the process is long-memory. Thus, spectral values at all discrete Fourier frequencies can be used in the optimisation for finding the debiased Whittle estimates, in contrast to the WLE from Section 3.2. Moreover, we can compute the expected periodogram exactly using Equation (3.3.4), in contrast to the WLE where the spectrum must be approximated as discussed in Section 3.2. Therefore, the DWLE has a lot of advantages over WLE for FGN with regards to efficient implementation, as we

shall discuss further.

3.4 Log-Periodogram Regression Estimator

The exact power spectral density function of FGN consists of an infinite sum, so either its approximated aliased spectrum or its unaliased spectrum is used for the WLE, as already discussed in Section 3.2. The latter takes the form of

$$S^{(U)}(\omega) = 4\sin^2\left(\frac{\omega}{2}\right)\frac{A^2}{|\omega|^{2H+1}}, \qquad -\pi \leq \omega \leq \pi. \tag{3.4.1}$$

For low frequencies, $\sin(\omega/2) \approx \omega/2$. This simplifies the unaliased spectrum as

$$S^{(U)}(\omega) \approx A^2|\omega|^{1-2H}, \qquad \omega \approx 0. \tag{3.4.2}$$

Then the logarithm is applied to both sides of Equation (3.4.2) such that

$$\log S^{(U)}(\omega) \approx (1-2H)\log|\omega| + 2\log(A), \qquad \omega \approx 0, \tag{3.4.3}$$

where A is independent of ω. If the periodogram is used as the spectral estimator of this unaliased spectrum, the quantity $1-2H$ can then be estimated from the slope of the linear regression line between the logarithm of the periodogram and the logarithm of ω at low frequencies. We choose $0 < \omega \leq \pi/8$ radians per unit time as our preference for these low frequencies. The estimator from this slope is named as the log-periodogram regression estimator (LPRE) in this thesis. For a stationary FGN, the slope is likely to be between -1 and 1. The estimation of H from this slope for FGNs with different true values of the Hurst exponent will be shown in Chapter 4.

3.5 Estimators for the Stationary FD Process

The stationary FD process is similar to FGN as they are both fractional processes with a single parameter controlling their long-memory behaviour. Although the Hurst exponent (H) and the order of fractional differencing (d) are obtained from two different processes, they are related by $H \approx d + 1/2$. To investigate such a relationship, in Chapter 4, we shall fit the time series of FGN with a stationary FD process. Both parameters in the stationary FD process, d and σ_ϵ^2, are estimated with the WLE and the DWLE using its own autocovariance sequence and power spectral density function in Equations (2.4.4) and (2.4.6), respectively.

3.6 Estimators for Discretely-Sampled FBM

The Hurst exponent (H) and the level variable (A) are two parameters in discretely-sampled FBM, denoted as $B = \{B_t\}_{t=1}^n$. To perform parameter estimation, one option is to take the first-order difference of this process and then fit FGN using the estimation methods described in Sections 3.1–3.4 to recover estimates of H and A. However, in Chapter 4, we will also study fitting FBM directly to data without taking this first-order difference. With the WLE, we fit FBM using three different forms of spectrum in the same way as fitting FGN in Section 3.2, but without using the squared gain function of the difference filter. These forms of spectrum are derived

from the following equation:

$$S_B(\omega) \approx \left(\frac{A^2}{(2\pi)^{2H+1}}\right)\left(\sum_{j=-M}^{M}\frac{1}{|f+j|^{2H+1}}+\right.$$
$$\sum_{l=-1,1}\left[\frac{1}{2H(lf+M+1)^{2H}}-\frac{(2H+1)(2H+2)(2H+3)}{720(lf+M+1)^{2H+4}}+\right.$$
$$\left.\left.\frac{2H+1}{12(lf+M+1)^{2H+2}}+\frac{1}{2(lf+M+1)^{2H+1}}\right]\right), \qquad -\pi \le \omega \le \pi. \quad (3.6.1)$$

The DWLE uses the expected periodogram instead of the power spectral density function. The exact equation of the expected periodogram of FBM cannot be derived due to its nonstationarity. However, with the adaptation of Equation (3.3.4) and the assumption of mean centering of the process, i.e., $\mathbb{E}[B_t] = 0$, the expected periodogram of FBM with discrete sampling can be approximated as

$$\begin{aligned}\bar{S}_B(\omega|\boldsymbol{\theta}) &\approx \sum_{\tau=-(n-1)}^{n-1}\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n-|\tau|}B_tB_{t+|\tau|}\right]e^{-i\omega\tau}\\ &= \sum_{\tau=-(n-1)}^{n-1}\left(\frac{V_HA^2}{2n}\sum_{t=1}^{n-|\tau|}(|t+\tau|^{2H}+|t|^{2H}-|\tau|^{2H})\right)e^{-i\omega\tau}\\ &= 2\cdot\mathrm{Re}\left(\sum_{\tau=0}^{n-1}\left(\frac{V_HA^2}{2n}\sum_{t=1}^{n-\tau}(|t+\tau|^{2H}+|t|^{2H}-|\tau|^{2H})\right)e^{-i\omega\tau}\right)\\ &\quad -\frac{V_HA^2}{n}\sum_{t=1}^{n}|t|^{2H}, \qquad -\pi \le \omega \le \pi. \end{aligned} \quad (3.6.2)$$

The time-averaged unaliased spectrum of FBM is $A^2|\omega|^{-1-2H}$, which is in a power-law form. For the LPRE, this relationship shows that the slope of the linear regression line is therefore $-1-2H$. Because H is between 0 and 1, the slope of the log-periodogram is likely to be between -3 and -1. This range of slope is completely different from that for FGN, which is between -1 and 1, and both ranges are divided by the boundary slope of -1. Thus, the LPRE can be used to evaluate whether or not the process is stationary using this slope.

3.7 Data Tapering

The periodogram in Equation 3.2.2 can be rewritten as

$$I(\omega) = \left| \sum_{t=1}^{n} \frac{X_t}{\sqrt{n}} \exp(-i\omega t) \right|^2, \qquad \omega \in \mathbf{\Omega}. \tag{3.7.1}$$

The multiplication of the time series and a finite sequence of $1/\sqrt{n}$, which we shall call the rectangular window, corresponds to a convolution in the frequency domain. Specifically, the frequency response of this rectangular window is the Dirichlet kernel, which is related to the expected periodogram via a convolution of the *squared* Dirichlet kernel with the spectrum (see Equation (3.3.2)). The squared Dirichlet kernel, also known as the Fejér kernel, consists of a main lobe centred at zero and sidelobes. The closed-form expression of the Fejér kernel has already been expressed in Equation (3.3.3). The number of its sidelobes increases as the length of time series or n increases. In addition, the highest value of kernel always occurs at the center of main lobe and this is equal to n. We illustrate a plot of several Fejér kernels with different values of n in Figure 3.7.1, as an example. The sidelobes of the Fejér kernel create an effect called "spectral leakage" resulting in the incorrect distribution of power across all frequencies. For FBM and other processes with a high dynamic range in the spectrum, this phenomenon causes a large discrepancy between its power spectral density function and the expected periodogram. This results in large bias of parameter estimates from the WLE.

Data tapering can be used to reduce such bias and discrepancy by suppressing these sidelobes. The tapered periodogram of $\{B_t\}$, which is a discretely-sampled

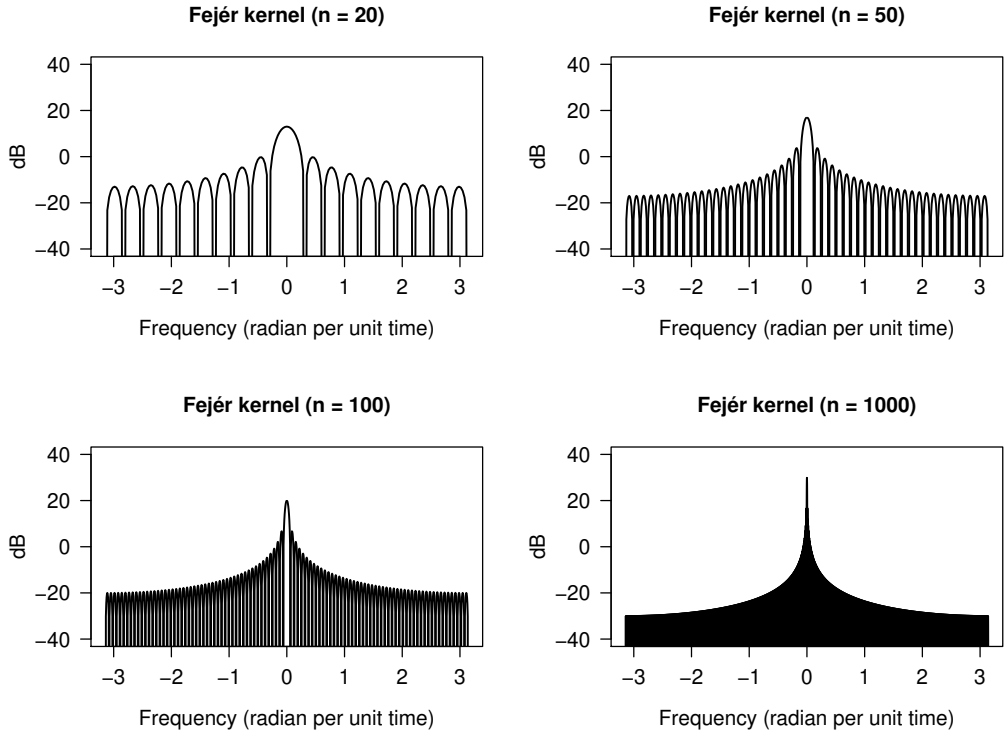

Figure 3.7.1: Fejér kernels with $n = \{20, 50, 100, 1000\}$.

FBM with length n, takes the form

$$J(\omega) = \left| \sum_{t=1}^{n} h_t B_t \exp(-i\omega t) \right|^2, \qquad \omega \in \mathbf{\Omega}, \tag{3.7.2}$$

where h_t is called the data taper such that $\sum_{t=1}^{n} {h_t}^2 = 1$. The general form of the periodogram, $I(\omega)$, utilises a rectangular window with the amplitude of $h_t = 1/\sqrt{n}$ for all $t = 1, 2, \ldots, n$ as its data taper. We shall refer to this choice of taper as the periodogram without tapering. However, when we use the periodogram with tapering,

we shall apply a 20% cosine taper that is given by

$$h_t = \begin{cases} \frac{C}{2}\left[1 - \cos\left(\frac{2\pi t}{\lfloor pn \rfloor + 1}\right)\right], & 1 \leq t \leq \frac{\lfloor pn \rfloor}{2}; \\ C, & \frac{\lfloor pn \rfloor}{2} < t < n + 1 - \frac{\lfloor pn \rfloor}{2}; \\ \frac{C}{2}\left[1 - \cos\left(\frac{2\pi(n+1-t)}{\lfloor pn \rfloor + 1}\right)\right], & n + 1 - \frac{\lfloor pn \rfloor}{2} \leq t \leq n, \end{cases} \tag{3.7.3}$$

where C is a normalising constant such that $\sum_{t=1}^{n} {h_t}^2 = 1$ as required, and p is set to 0.2 for this $p \times 100\%$ cosine taper. Figure 3.7.2 shows the difference between the rectangular and the 20% cosine tapers.

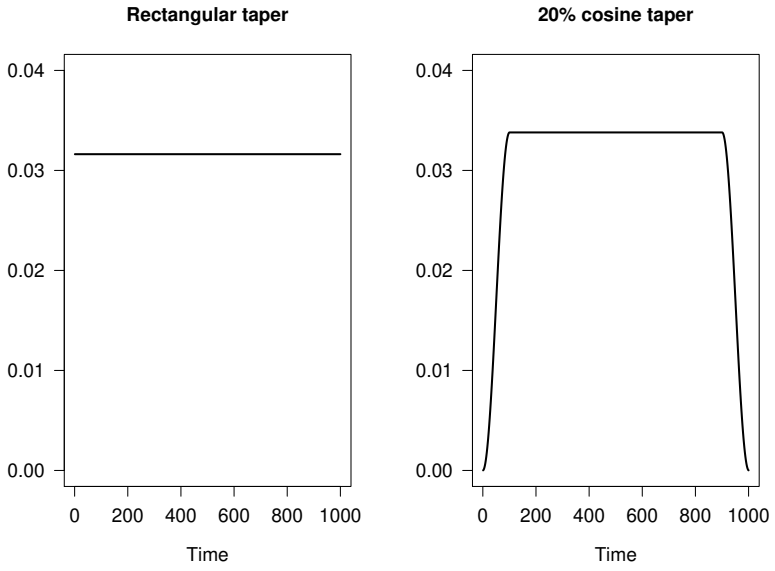

Figure 3.7.2: Rectangular (left) and 20% cosine (right) tapers.

We can also use the data taper in combination with the DWLE. The expected periodogram of discretely-sampled FBM with 20% cosine tapering is given by

$$\bar{S}_B^{(T)}(\omega|\boldsymbol{\theta}) = 2 \cdot \mathrm{Re}\left(\sum_{\tau=0}^{n-1} s_{B,(t,\tau)} h_t h_{t+\tau} e^{-i\omega\tau}\right) - \sum_{t=1}^{n} h_t^2 |t|^{2H}, \tag{3.7.4}$$

where $s_{B,(t,\tau)}$ is a time-dependent autocovariance sequence of discretely-sampled FBM, which is defined as

$$s_{B,(t,\tau)} = \frac{V_H A^2}{2} \sum_{t=1}^{n-|\tau|} (|t+\tau|^{2H} + |t|^{2H} - |\tau|^{2H}). \tag{3.7.5}$$

These forms of the expected periodogram (Equations (3.6.2) and (3.7.4)) are expensive to compute in the DWLE versus those for FGN (Equation (3.3.4)), but in the next section, we shall explore the effectiveness of each estimation method in practice under wide-ranging simulation studies.

Chapter 4

Simulation and Empirical Studies

4.1 Preliminary Analysis

The simulation of discrete-time FGN was implemented in RStudio with R version 4.0.0. First, we generated the autocovariance matrix of FGN, which is a Toeplitz matrix derived from the autocovariance sequence with a true value of the Hurst exponent, denoted as H_T, and a true value of the variance, denoted as σ_T^2. Each FGN was simulated by the multiplication of the lower triangular matrix from the Cholesky decomposition of this autocovariance matrix and a vector of random numbers from the Gaussian distribution with zero mean and unit variance. We used this technique for all simulated FGNs with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\sigma_T^2 = 1$, and several examples of these processes with different H_T were already shown in Figure 2.3.1 in Chapter 2. The cumulative sum of FGN results in discretely-sampled FBM, and several examples of these were also shown in Figure 2.2.1.

All types of possible spectra used in Whittle likelihood estimation will now be

compared. These include all three different forms of spectrum and the expected periodogram, which were introduced earlier in Sections 3.2 and 3.3. These spectra of simulated FGNs with $\sigma_T^2 = 1$ are illustrated in Figure 4.1.1 for $H_T \in \{0.1, 0.3, 0.5\}$ and Figure 4.1.2 for $H_T \in \{0.7, 0.9\}$.

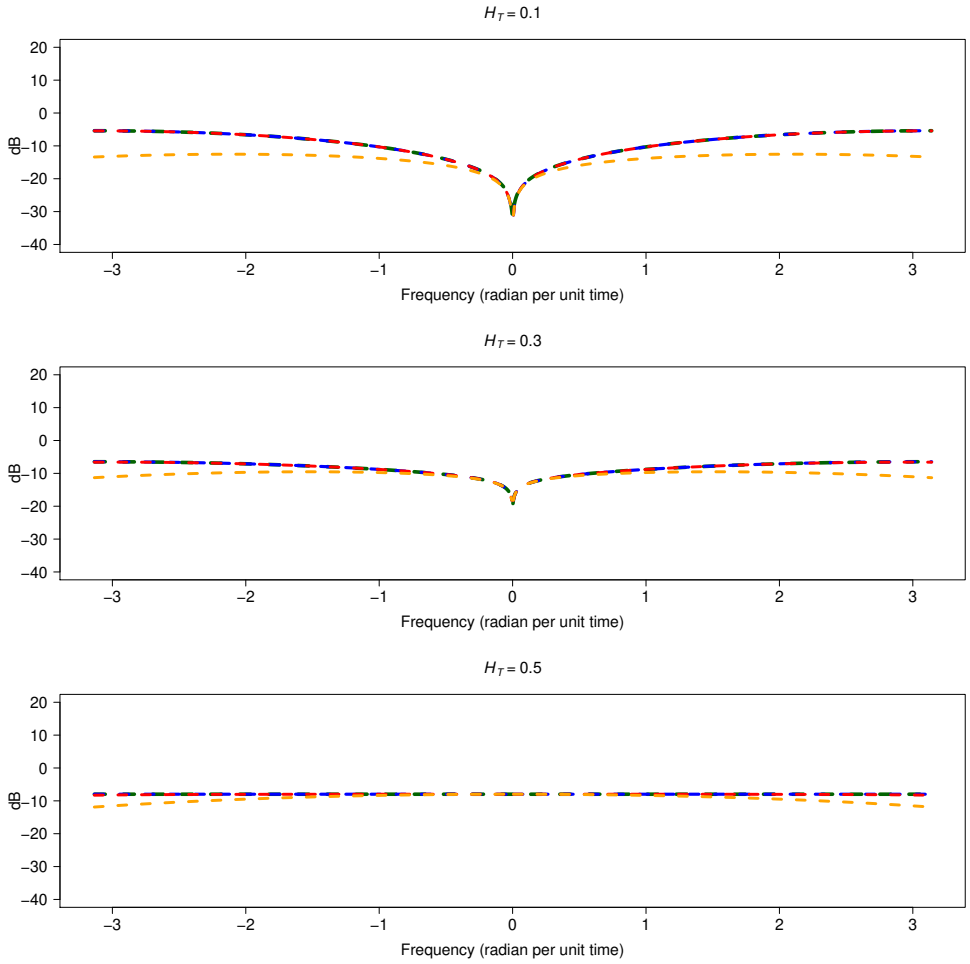

Figure 4.1.1: The expected periodogram (green), the approximated aliased power spectral density functions with $M = 100$ (blue) and $M = 0$ (red), and the unaliased power spectral density function (orange) of each simulated discrete-time FGN with $H_T \in \{0.1, 0.3, 0.5\}$ and $\sigma_T^2 = 1$. The green, blue and red lines are often overlaid but small differences do exist.

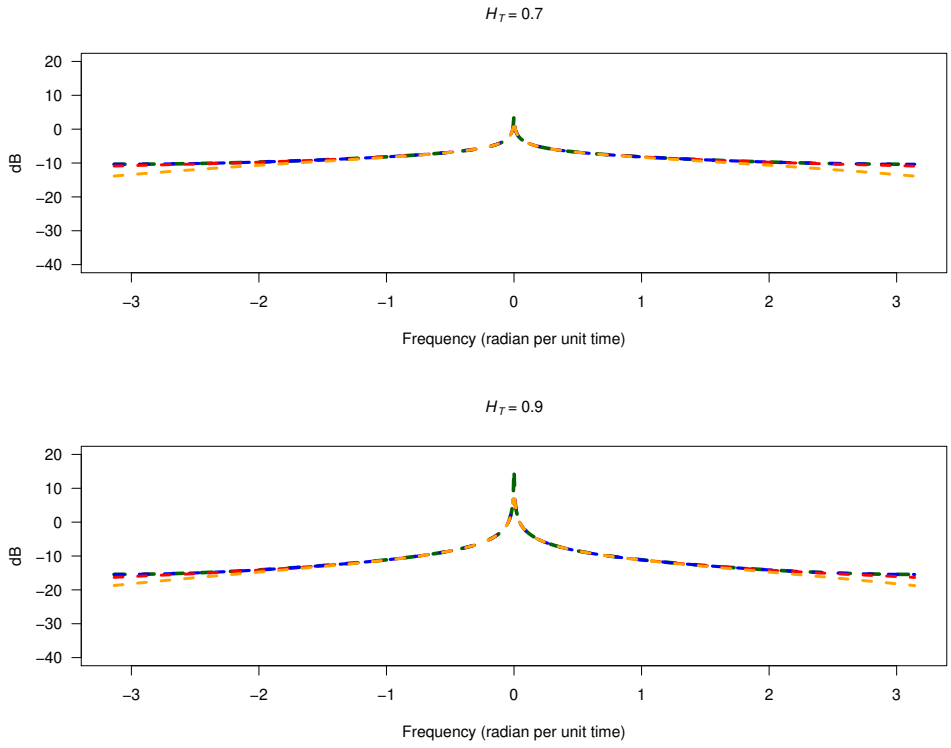

Figure 4.1.2: The expected periodogram (green), the approximated aliased power spectral density functions with $M = 100$ (blue) and $M = 0$ (red), and the unaliased power spectral density function (orange) of each simulated discrete-time FGN with $H_T \in \{0.7, 0.9\}$ and $\sigma_T^2 = 1$. The green, blue and red lines are often overlaid but small differences do exist.

We observe that the expected periodogram and the aliased spectra of each FGN in these two figures are very close at all frequencies as green, blue, and red lines are almost inseparable in each plot. The unaliased spectrum is similar to the aliased spectrum with $M = 0$ except that the summation terms including the parameter l in Equation (2.3.7) are all removed. This results in its lower spectral values than other spectra in the figure, especially evident at high frequencies. According to this finding,

we expect a noticeable difference of estimation results of parameters for FGN with the WLE using each of these spectral forms, as will be discussed later.

Unlike FGN, the spectrum of FBM is always the highest at the zero frequency for all values of H_T. All spectra and the expected periodogram (without tapering) of discretely-sampled FBM with $A = 1$ are shown in Figure 4.1.3 for $H_T \in \{0.1, 0.3, 0.5\}$ and 4.1.4 for $H_T \in \{0.7, 0.9\}$.

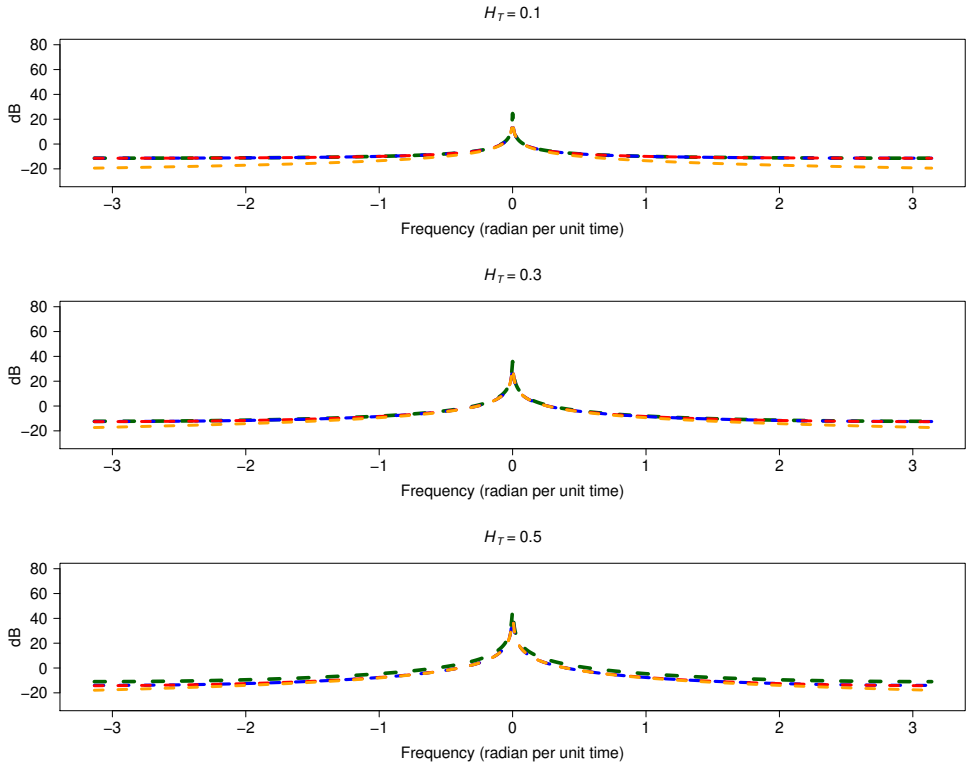

Figure 4.1.3: The expected periodogram without tapering (green), the approximated aliased power spectral density functions with $M = 100$ (blue) and $M = 0$ (red), and the unaliased power spectral density function (orange) of each simulated discretely-sampled FBM with $H_T \in \{0.1, 0.3, 0.5\}$. The level parameter of FBM is set as one, i.e., $A = 1$.

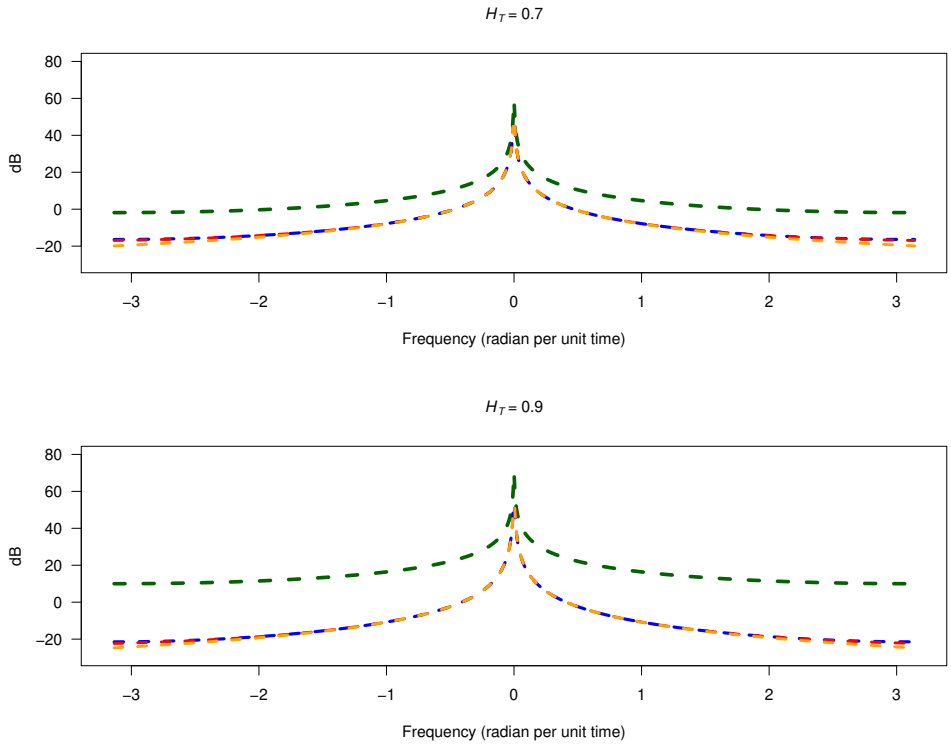

Figure 4.1.4: The expected periodogram without tapering (green), the approximated aliased power spectral density functions with $M = 100$ (blue) and $M = 0$ (red), and the unaliased power spectral density function (orange) of each simulated discretely-sampled FBM with $H_T \in \{0.7, 0.9\}$. The level parameter of FBM is set as one, i.e., $A = 1$.

From these two figures, the relationship between all three different forms of spectrum of FBMs is similar to those of FGNs such that the unaliased spectrum is lower than the others. The expected periodogram without tapering is noticeably higher than all other functions especially when $H_T \geq 0.5$, and this is due to accounting for spectral leakage from the Fejér kernel. The 20% cosine taper could be applied to reduce the large bias of the spectral estimator of the expected periodogram. Figure 4.1.5

shows that this taper can efficiently reduce the level of the expected periodogram to be almost the same as the aliased spectrum for all cases of H_T.

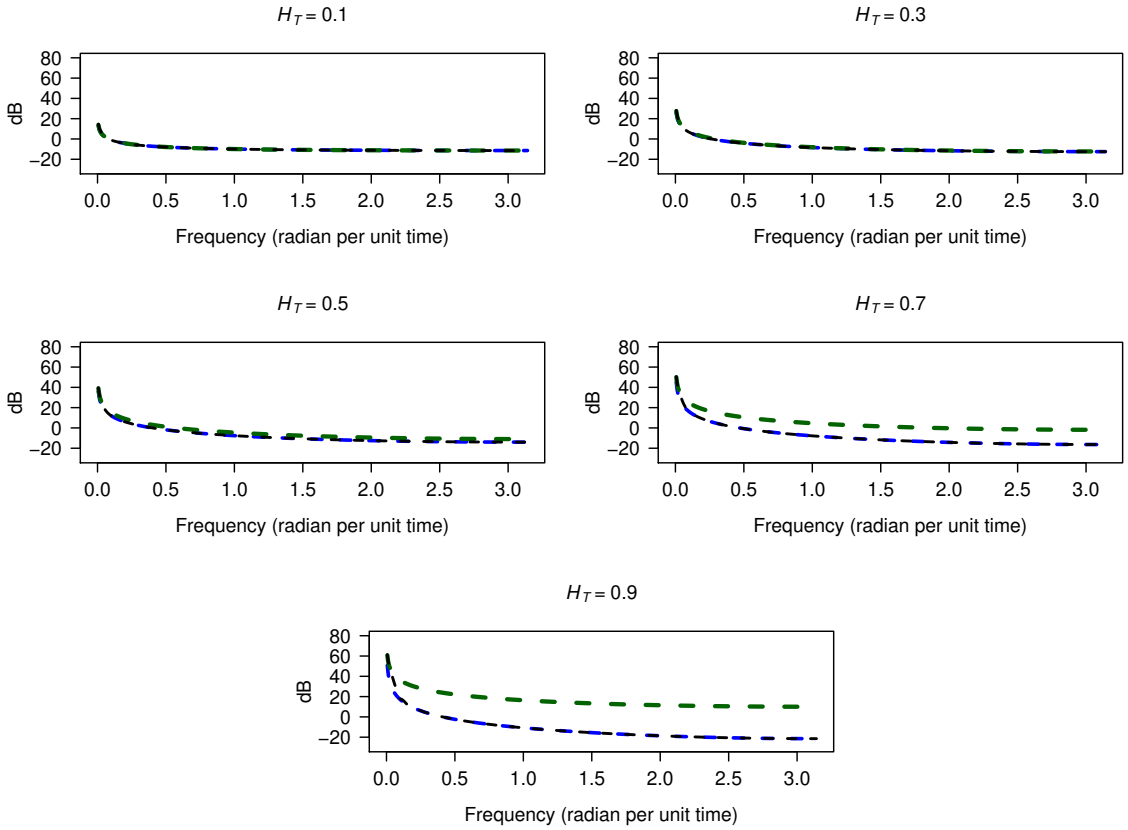

Figure 4.1.5: The approximated aliased power spectral density function with $M = 100$ (blue), and the expected periodograms without (green) and with 20% cosine tapering (black) of each simulated FBM with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $A = 1$.

There exists an asymptotic property of the periodogram such that twice the ratio of the periodogram to the spectral density function is chi-squared distributed with two degrees of freedom, i.e., $2(I(\omega)/S(\omega|\boldsymbol{\theta})) \sim \chi^2_2$ as $n \to \infty$, where $S(\omega|\boldsymbol{\theta})$ is the spectrum of either FGN or FBM. This distribution is equivalent to the exponential distribution with the rate parameter of 1/2. Figure 4.1.6 presents the Q-Q plots of

such a distribution derived from simulated FGN and FBM with long memory, both with $H_T = 0.7$ and time series of length $n = 1000$. Their exact spectrum is replaced by its aliased form with $M = 100$, whereas the periodogram of FBM including the 20% cosine taper, as given by $J(\omega)$ in Equation (3.7.2), is used instead of $I(\omega)$. From this figure, both plots show that these measures from simulated fractional processes are highly likely to be chi-squared distributed with two degrees of freedom as most points lie on the straight diagonal line with $y = x$. A few points deviate from this line and this is due to the finite length of our simulated time series. These properties were also found from simulated fractional processes with different values of H_T or different lengths of time series.

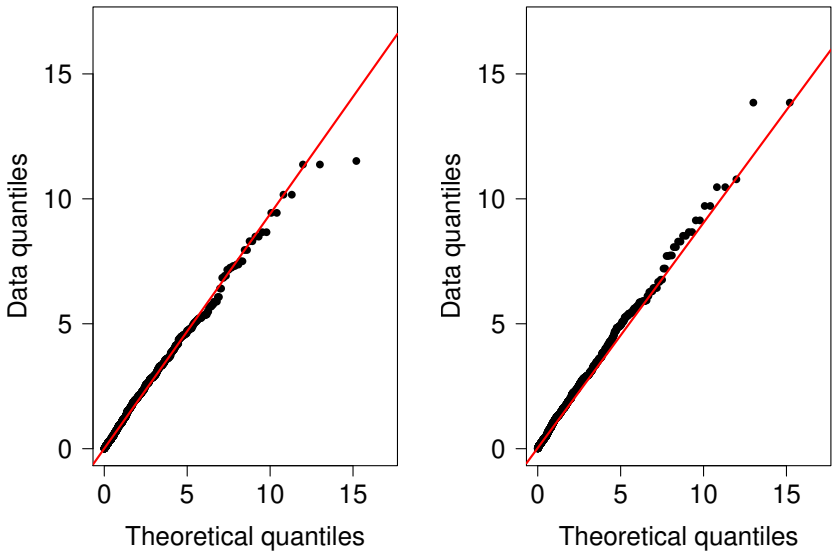

Figure 4.1.6: The chi-squared Q-Q plots of the distribution between the chi-squared distributed data with two degrees of freedom, and either FGN (left) or FBM (right) with $H_T = 0.7$, and $\sigma_T^2 = 1$ (FGN) or $A = 1$ (FBM).

Figure 4.1.7 shows several examples of linear regression of the periodogram against ω in log-log space at low frequencies from simulated FGNs and FBMs with $H_T = \{0.3, 0.5, 0.7\}$, and $\sigma_T^2 = 1$ (FGNs) or $A = 1$ (FBMs). The estimated Hurst exponents derived from the slopes of these regression lines are 0.327, 0.469, and 0.726, respectively, for simulated FGNs, and 0.255, 0.454, and 0.697, respectively, for simulated FBMs. A regression line of a fractional process with long-memory time series always has a negative slope. The comparison of estimation results between this method and all other parametric estimators will be discussed later in this chapter.

4.2 Estimation Results from Simulated FGNs

First, we simulated five sets of zero-mean FGNs. Each set contains 10000 time series of this fractional process with the same length $n = 1000$, each with the same true value of the Hurst exponent with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the same true value of the sample variance with $\sigma_T^2 = 1$. The order of the simulated dataset corresponds to the values of H_T in ascending order, for example, the first dataset has $H_T = 0.1$, and the fifth dataset has $H_T = 0.9$. The Hurst exponent and the sample variance of each simulated FGN were simultaneously estimated with each likelihood-based method given in Chapter 3 (excluding the MLE) using a derivative-free search called the Hooke-Jeeves method (Hooke and Jeeves, 1961). This optimisation method finds an optimal minimum point of the bivariate function which is the negative of log-likelihood function with constraints of $0 < H < 1$ and $\sigma_X^2 > 0$. This method was implemented by the "fminsearch" function in the package "pracma" (Borchers, 2022)

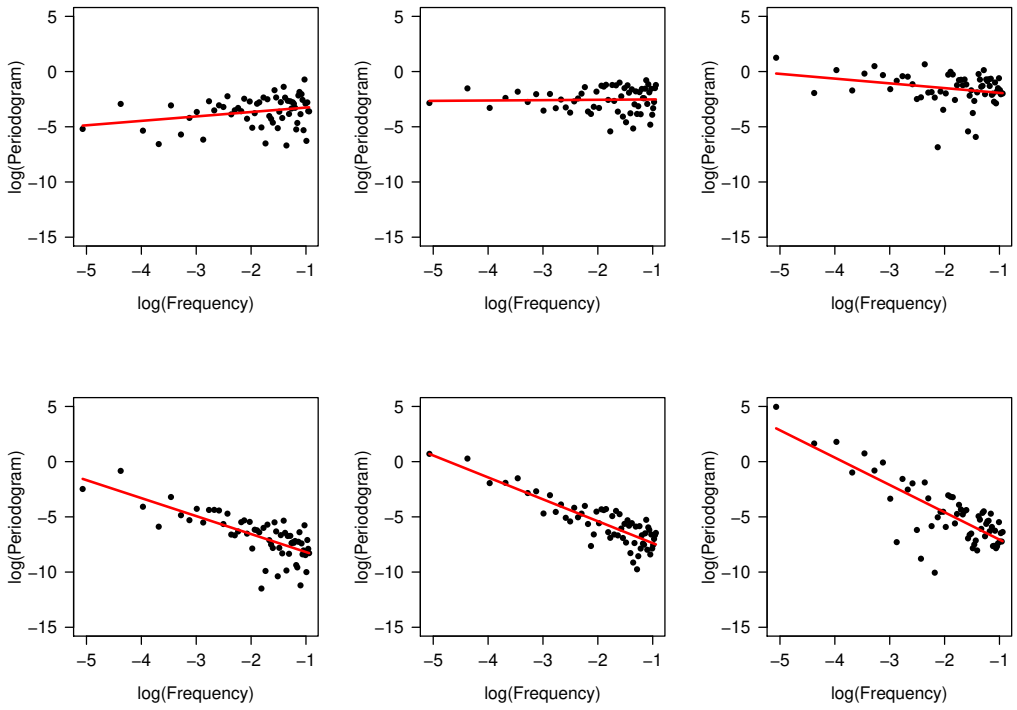

Figure 4.1.7: The linear regression plots for the estimation of the Hurst exponent from simulated FGNs (upper) and FBMs (lower), each with $H_T = 0.3$ (left), $H_T = 0.5$ (middle), and $H_T = 0.7$ (right). In each plot, the x-axis is the low frequencies between 0 and $\pi/8$ radians per unit time, and the y-axis is either the periodogram ($I(\omega)$) of FGN or the periodogram with 20% cosine tapering ($J(\omega)$) of FBM. Both axes are in logarithmic scales.

in RStudio. All estimation methods for simulated FGNs are numerically listed as follows:

1. WLE using the approximated aliased spectrum with $M = 100$ (Equation (3.2.3)).

2. WLE using the approximated aliased spectrum with $M = 0$ (Equation (3.2.3)).

3. WLE using the unaliased spectrum (Equation (3.2.3)).

4. DWLE (Equation (3.3.5)).

5. LPRE (Equation (3.4.3)).

We remind the reader that all forms of spectrum of FGN used for Methods 1, 2, and 3 are given in Equations (2.3.6) and (2.3.7). The mean and standard deviation of estimated values of the Hurst exponent and the sample variance from each set of simulated FGN are summarised in Tables 4.2.1 and 4.2.2, respectively.

Table 4.2.1: The mean and standard deviation (mean ± SD) of estimated values of the Hurst exponent (H) from each of the five different sets of FGNs.

Method	First set ($H_T = 0.1$)	Second set ($H_T = 0.3$)	Third set ($H_T = 0.5$)	Fourth set ($H_T = 0.7$)	Fifth set ($H_T = 0.9$)
1	0.102 ± 0.012	0.299 ± 0.017	0.499 ± 0.020	0.699 ± 0.021	0.901 ± 0.022
2	0.102 ± 0.012	0.297 ± 0.017	0.494 ± 0.019	0.691 ± 0.020	0.888 ± 0.021
3	0.005 ± 0.001	0.163 ± 0.024	0.399 ± 0.023	0.619 ± 0.023	0.833 ± 0.023
4	0.099 ± 0.012	0.299 ± 0.017	0.499 ± 0.019	0.699 ± 0.020	0.899 ± 0.020
5	0.027 ± 0.064	0.288 ± 0.061	0.499 ± 0.062	0.705 ± 0.061	0.910 ± 0.061

It is clear that Method 3 provides the worst estimation results with the largest bias of both H and σ_X^2 since this method does not take into account aliasing from the discrete sampling of FGN. We can estimate σ_X^2 through the parameter A, as shown in

Table 4.2.2: The mean and standard deviation (mean ± SD) of estimated values of the sample variance with $\sigma_T^2 = 1$ from each of the five different sets of FGNs.

Method	First set ($H_T = 0.1$)	Second set ($H_T = 0.3$)	Third set ($H_T = 0.5$)	Fourth set ($H_T = 0.7$)	Fifth set ($H_T = 0.9$)
1	0.997 ± 0.050	0.999 ± 0.046	1.000 ± 0.044	1.002 ± 0.056	1.064 ± 0.253
2	1.000 ± 0.050	1.007 ± 0.047	1.011 ± 0.045	1.008 ± 0.055	0.977 ± 0.186
3	50.896 ± 2.477	2.616 ± 0.370	1.490 ± 0.082	1.173 ± 0.054	0.824 ± 0.096
4	0.999 ± 0.050	0.999 ± 0.046	1.000 ± 0.044	1.000 ± 0.055	1.028 ± 0.206

Equation (2.3.3) by estimating the intercept term in the linear regression. However, this is likely to provide high bias as Equation (3.4.3) is effective when ω is very close to zero. Thus, Method 5 was only used to estimate the Hurst exponent in this analysis. This method works quite well on FGN with H close to 0.5, but a large bias is observed when H is close to 0 or 1. This is due to the weak correlation from the regression plot of the logarithm of the periodogram, and the possibility that an estimated value of the Hurst exponent from the slope is outside the range of H between 0 and 1. These issues lead to a drawback of the LPRE. However, an important benefit of this estimator is that it does not require any optimisation method for the estimation such that it requires less computational time than other estimators proposed in this thesis.

Next, we show violin plots of estimated values of H and σ_X^2 from simulated FGNs

with $H_T = 0.1$ (highly anti-persistent short-memory process) and $H_T = 0.9$ (highly persistent long-memory process) in Figures 4.2.1 and 4.2.2. We exclude Method 3 due to its very poor performance across all values of H_T. The violin plots from the LPRE (Method 5) show higher bias and variance than the others, and this agrees with the results shown in Table 4.2.1. The violin plots from Methods 1, 2, and 4 in Figure 4.2.1 are nearly the same. Their estimated values of H are approximately normally distributed. Their mean values are all close to H_T, and their variances are almost identical. In Figure 4.2.2, the violin plots from these three methods are also very similar. However, for very high H_T, there is a possibility that the estimated sample variance is much larger than its true value. This causes very high variance of estimated values of sample variance shown by both the violin plot and Table 4.2.2. One way of explaining this overestimation is that a simulated FGN with high H_T sometimes has very similar fractional behaviour to a nonstationary FBM with low H_T and non-constant variance. This can result in high estimated value of the sample variance, caused by the persistent nature of the time series sampling larger than expected values repetitively (see Figure 2.3.1 with $H_T = 0.9$ and compare this with Figure 2.2.1 with $H_T = 0.1$). Despite this problem, we found that the mean of estimated values of the sample variance with large sample size is close to its true value and its distribution is still bell-shaped with slightly positive skewness.

To further investigate the impact of H on estimation performance, we now report on two metrics for measuring estimation errors (or measurement errors). These include the mean absolute error (MAE) and the root-mean-square deviation (RMSD). Here we simulated FGNs with a higher range and resolution of values of H_T such that

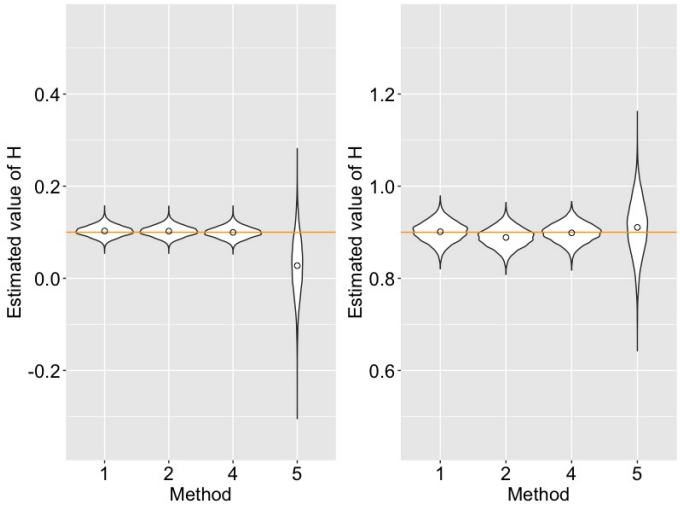

Figure 4.2.1: The violin plots of estimated values of the Hurst exponent from simulated FGNs with $H_T = 0.1$ (left) and $H_T = 0.9$ (right) using four different estimation methods. The orange horizontal line indicates H_T of these simulated processes.

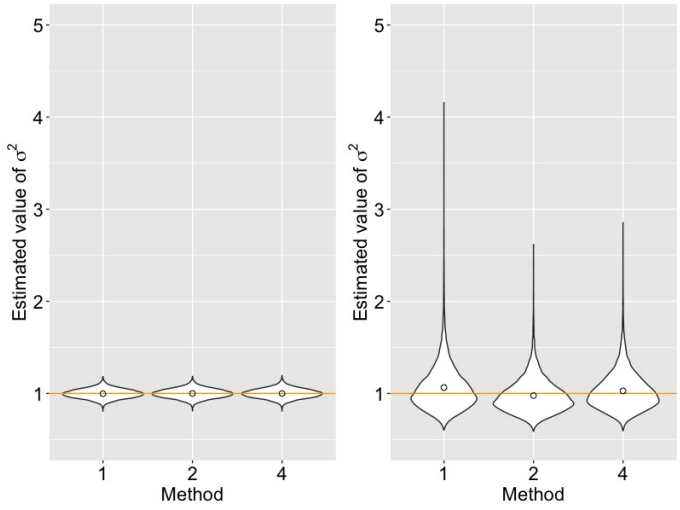

Figure 4.2.2: The violin plots of estimated values of the sample variance from simulated FGNs with $H_T = 0.1$ (left) and $H_T = 0.9$ (right) using three different estimation methods. The orange horizontal line indicates σ_T^2 of these simulated processes ($\sigma_T^2 = 1$).

$H_T = 0.05 \times S$, where S is a positive number from 1 to 19, and $\sigma_T^2 = 1$. The length of time series of each process is $n = 1000$, and the number of simulated time series is set as 1000 for each value of H_T. Figure 4.2.3 shows the measurement errors of estimates derived from Methods 1, 2, and 4. Methods 3 and 5 are excluded from this plot due to their large bias of estimates, which has already been discussed. From this figure, all three estimators have very close measurement errors when $H_T \leq 0.5$. However, when $H_T > 0.5$, their measurement errors are quite different especially those derived from time series of simulated FGNs with H_T close to one. The DWLE provides the lowest measurement errors of both estimates. The WLE with $M = 100$ and with $M = 0$ perform similarly with the former giving lower measurement errors of the Hurst exponent, and the latter giving lower measurement errors of the sample variance. Overall, we would expect $M = 100$ to be the better choice of the two as it uses a more accurate approximation of the power spectral density function, however it does take substantially longer to compute (as we shall shortly discuss further). According to all plots in this figure, we may conclude that Method 4 or the DWLE is likely to be the most appropriate method for the simultaneous estimation of the Hurst exponent and the sample variance.

We propose two more reasons why the DWLE should be an optimal choice for the estimation of parameters of FGN. First, this estimator utilises all discrete Fourier frequencies in its optimisation of the log-likelihood function whether or not the time series is long-memory. The WLE, however, excludes the spectral value at the zero frequency as the spectrum approaches infinity at this frequency for the long-memory time series, and this will contribute to slightly larger error of estimates than the DWLE.

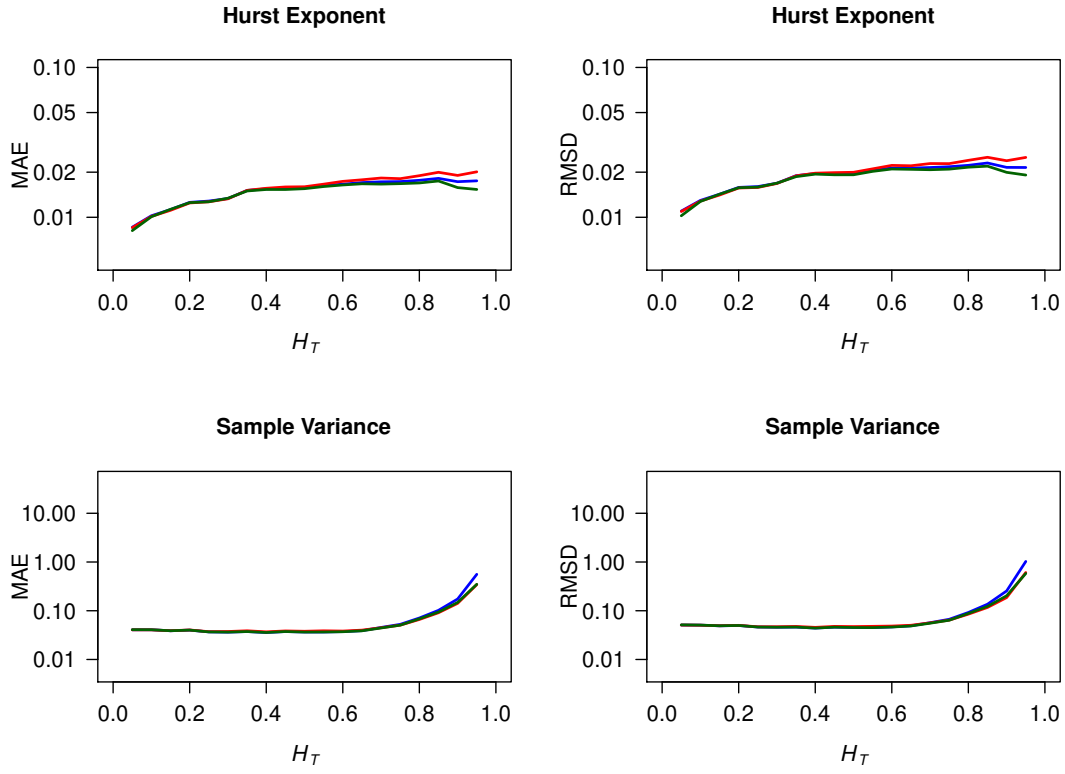

Figure 4.2.3: The line plots of mean absolute error and root-mean-square deviation from the simultaneous estimation of both parameters of simulated FGNs with $\sigma_T^2 = 1$ and H_T varied between zero and one. Three estimators are used for each plot including the WLE with $M = 100$ (blue), the WLE with $M = 0$ (red), and the DWLE (green). The y-axis of each plot is in the logarithmic scale.

Second, the DWLE requires less computational time or CPU time than the WLE for the simultaneous estimation of parameters of FGN with any value of H_T between zero and one, as shown in Figure 4.2.4. The computation time for the summation of the approximated aliased spectrum of FGN in Equation (2.3.7) is slightly slower then the computational time for the fast Fourier transform of the expected periodogram in Equation (3.3.4) at each iteration of the optimisation procedure in the "fminsearch"

function. However, a large number of iterations to simultaneously find the optimised values of the Hurst exponent and the sample variance result in much longer computational time for the WLE with the approximated aliased spectrum than the DWLE with the expected periodogram, especially for the spectrum with $M = 100$. In addition, there is a slight increase of the computational time when H_T approaches one for each estimation method. This is likely due to the overestimation problem of σ_X^2 from simulated FGN with high H_T, as explained earlier, which makes the optimisation take longer to converge as a wider range of σ_X^2 values are considered.

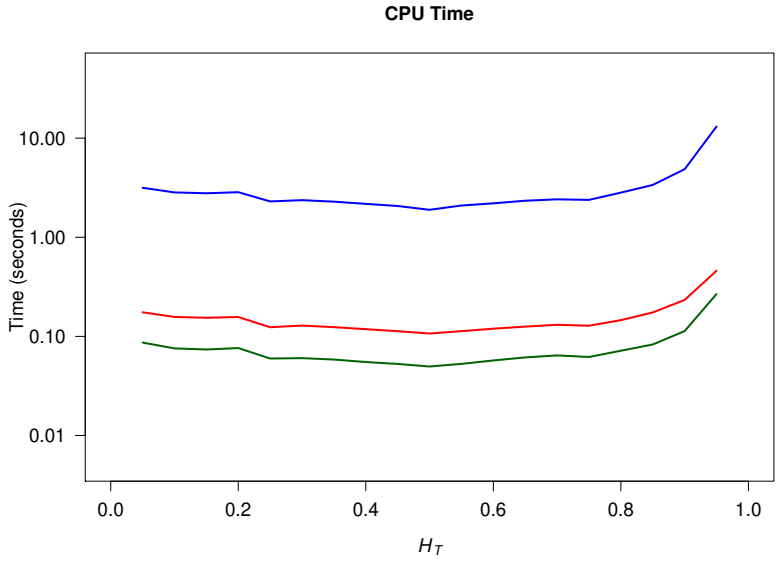

Figure 4.2.4: The line plots of average computation time or CPU time for the simultaneous estimation of both parameters of simulated FGNs with $\sigma_T^2 = 1$ and H_T varied between zero and one. Three estimators are used including the WLE with $M = 100$ (blue), the WLE with $M = 0$ (red), and the DWLE (green). The y-axis is in the logarithmic scale. CPU time is computed on a 1.8 GHz dual-core Intel Core i5 processor.

Next, we investigate the effect of different values of length n on the estimation of parameters of simulated FGNs using the DWLE. Each set of 10000 time series of FGNs with the same length $n \in \{256, 512, 1024, 2048, 4096\}$, the same $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the same $\sigma_T^2 = 1$ was simulated. Figures 4.2.5 and 4.2.6 show the line plots of mean absolute error (MAE), standard deviation (SD), root-mean-square deviation (RMSD), and average of CPU time derived from the simultaneous estimation of H and σ_X^2 against the logarithmic function of length n. All these measures except the average of CPU time are in the logarithmic scale. These two figures suggest that the MAE and the RMSD of both estimates decrease as n increases. They are likely to be power-law functions of length n due to a nearly constant slope of each line in their plots. Hence, we can expect that these measurement errors are likely to approach zero as n approaches infinity, and this shows the consistency of these estimates from the DWLE.

The logarithmic function of standard deviation of each estimate is also nearly linearly related to the logarithmic function of length n. According to Sykulski et al. (2019), the debiased Whittle likelihood estimates are $\sqrt{n}$-consistent, and this was proven with the assumption that the time series are short-memory. Although we have not established a complete mathematical proof for time series with long memory, we can show that this convergence rate is likely to be applicable with time series of long-memory process. Table 4.2.3 reports estimated slopes from regression lines of standard deviation of estimates on length n in log-log space. In Equation (3.1.5), we reported that estimates from the MLE are asymptotically efficient and $\sqrt{n}$-consistent. If we assume that this statement also holds for estimates from the DWLE, we can expect

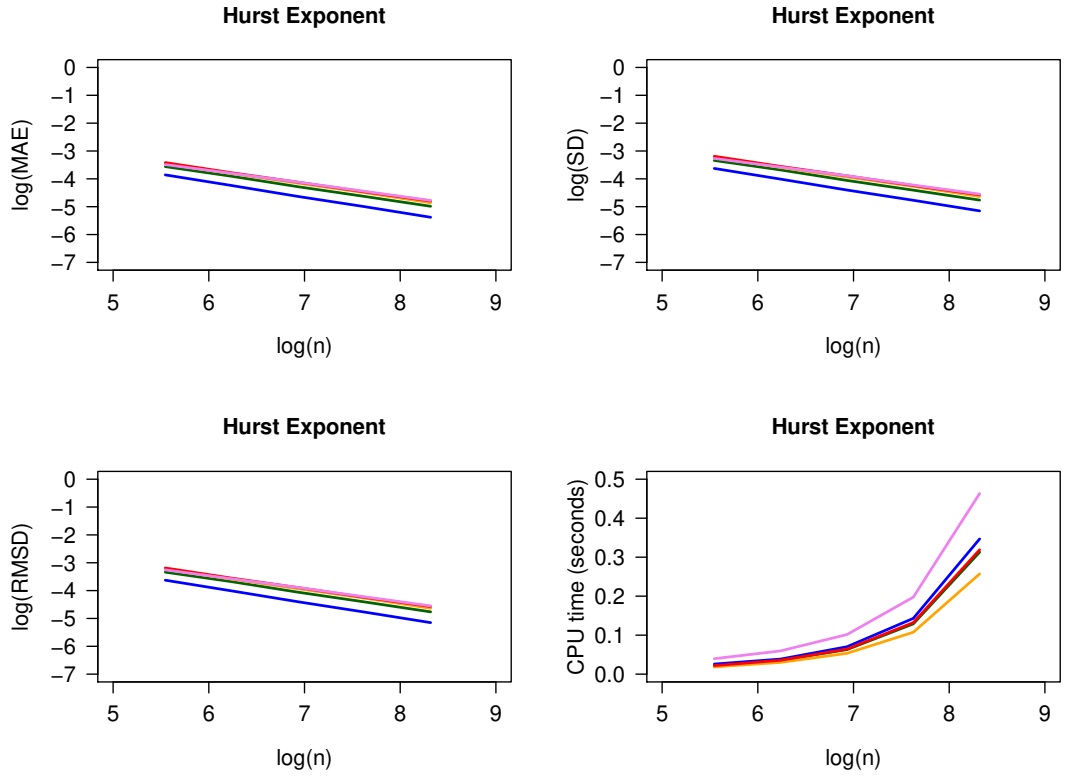

Figure 4.2.5: The line plots of the logarithmic functions of mean absolute error, standard deviation, root-mean-square deviation, and the linear function of average CPU time against the logarithmic function of length n from the estimation of the Hurst exponent of simulated FGNs with $H_T = 0.1$ (blue), $H_T = 0.3$ (green), $H_T = 0.5$ (orange), $H_T = 0.7$ (red), and $H_T = 0.9$ (violet) using the debiased Whittle likelihood estimator.

that the standard deviation of estimates is a power-law function of length n with an exponent of -0.5, and this number is expressed by a slope of the linear regression between standard deviation and length n in log-log space. Our estimated slopes are all close to this value of power exponent despite slightly deviations from simulated FGNs with very high H_T, for example, $H_T = 0.9$. We may deduce that the debiased Whittle

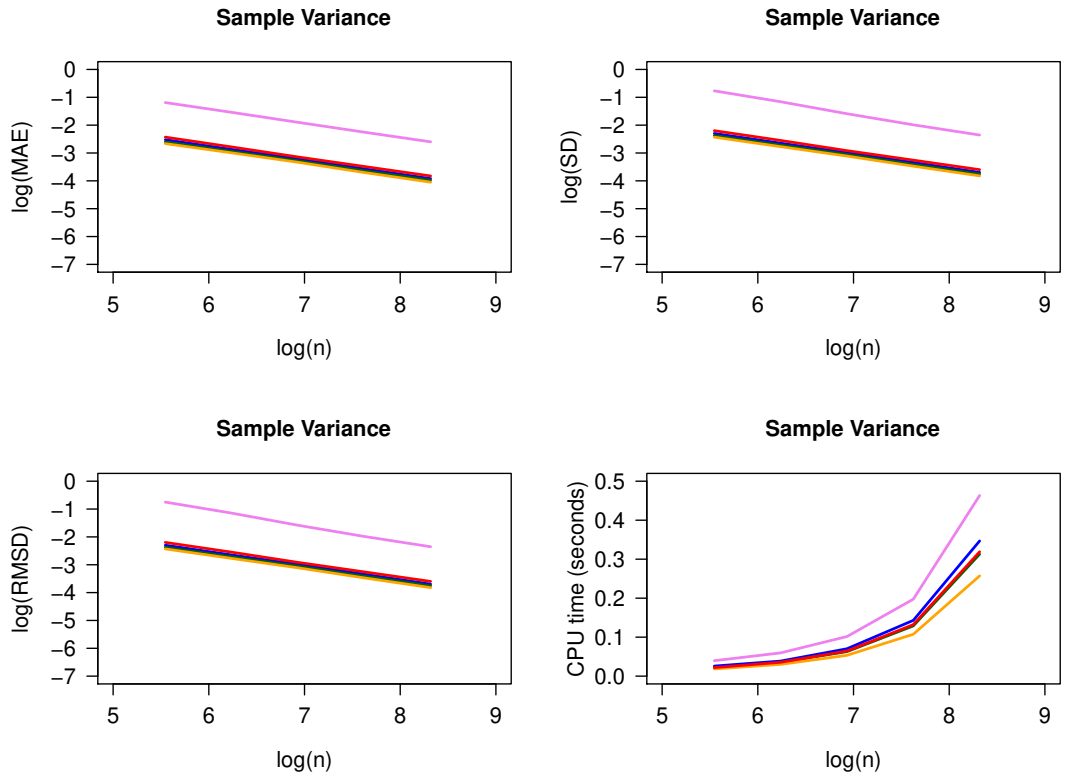

Figure 4.2.6: The line plots of logarithmic functions of mean absolute error, standard deviation, root-mean-square deviation, and the linear function of average CPU time against the logarithmic function of length n from the estimation of the sample variance of simulated FGNs with $H_T = 0.1$ (blue), $H_T = 0.3$ (green), $H_T = 0.5$ (orange), $H_T = 0.7$ (red), and $H_T = 0.9$ (violet) using the debiased Whittle likelihood estimator.

likelihood estimates are approximately $\sqrt{n}$-consistent and asymptotically efficient for FGN, according to our findings from Figures 4.2.5 and 4.2.6. This is consistent with findings in Robinson (2003), which state that the Whittle likelihood estimates are also $\sqrt{n}$-consistent for long-memory processes. In addition, these two figures consist of the plots of average CPU time, and these two plots are identical due to the simultaneous estimation of both parameters. Longer time series lengths require more average CPU

time than shorter lengths, as expected. In general this rate of increase is expected to scale as $\mathcal{O}(n \log n)$ as n increases, due to the use of FFTs in DWLE, but the precise relationship will depend on how the algorithms are implemented as n becomes very large, the study of which we reserve for future work.

Table 4.2.3: Estimated slopes from regression lines of standard deviation of estimates on length n in log-log space. These slopes are derived from each of the five different sets of FGNs with different values of H_T.

Parameter	$H_T = 0.1$	$H_T = 0.3$	$H_T = 0.5$	$H_T = 0.7$	$H_T = 0.9$
H	-0.549	-0.516	-0.501	-0.498	-0.463
σ_X^2	-0.504	-0.500	-0.502	-0.506	-0.577

In Chapter 2, we showed that the relationship between long-memory parameters of FGN and the stationary FD process is given by $H = d + 1/2$. This means that the memory property between FGN with the estimated parameter of $\hat{H}$ and the FD process with the estimated parameter of $\hat{d} = \hat{H} - 1/2$ is theoretically identical. This motivates us to apply the model of FD processes to simulated FGNs and observe how the parameter d is estimated for each simulated FGN comparing with its estimated value of the Hurst exponent. This also tests the ability of each estimation method to be robust to related but misspecified long-memory models. We used all the same simulated processes as those for analysing measurement errors in Figure 4.2.3. The measurement errors including the mean absolute error (MAE) and the root-mean-

square deviation (RMSD) from the estimation of d using the WLE with $M = 100$ and the DWLE are presented by the line plots in Figure 4.2.7. Other methods of the WLE are possible to use, however they provide larger measurement errors than the WLE with $M = 100$. This figure shows that both methods provide very similar measurement errors. These errors from the debiased method are slightly lower at most ranges of H_T except at both extremes of H_T. The MAE and the RMSD from the estimation of d are likely to be the lowest when the simulated FGN is very close to Gaussian white noise, i.e., $H_T = 0.5$, which is as expected as both processes will then exhibit very similar spectra and autocovariances.

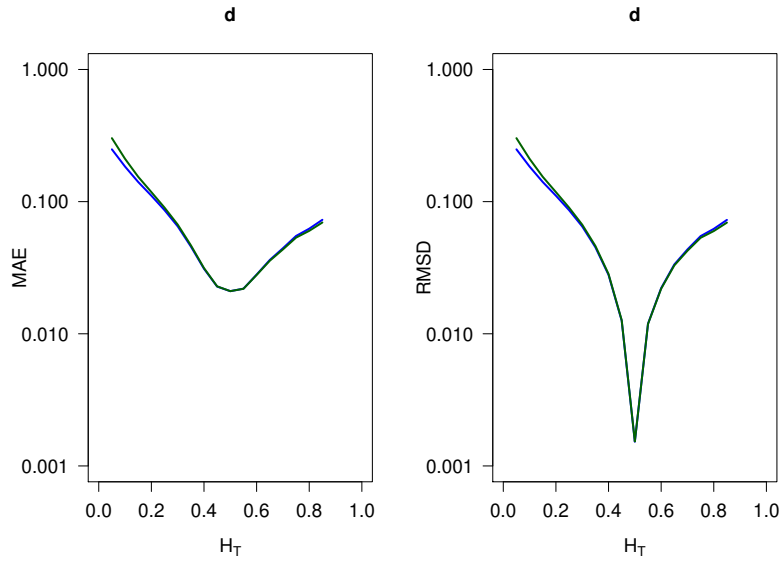

Figure 4.2.7: The line plots of mean absolute error and root-mean-square deviation from the estimation with the modelling of FD processes to the simulated FGNs used in Figure 4.2.3. Two estimators are used for each plot including the WLE with $M = 100$ (blue) and the DWLE (green). The y-axis of both plots is in the logarithmic scale.

Figure 4.2.8 shows the mean of the difference between estimates (the mean of

$\hat{d} - \hat{H}$) from each set of simulated FGNs with a fixed value of H_T varied between zero and one. This mean of difference corresponds to its theoretical aspect such that $\hat{d} - \hat{H} = -1/2$ only when $H_T = 0.5$. Using the approximation of $\sin(\omega/2) \approx \omega/2$, the power spectral density function of FD process in Equation (2.4.6) at low frequencies can be approximated as

$$S(\omega) = \sigma_\varepsilon^2 |\omega|^{-2d}, \qquad \omega \approx 0. \tag{4.2.1}$$

The variance of the white noise process or σ_ε^2 is only used to set the level of spectrum although we estimate this parameter along with the order of differencing for modelling of FD processes. We compare this equation with the unaliased power spectral density function of FGN at low frequencies derived from the approximation of $\sin(\omega/2) \approx \omega/2$ and the substitution of $j = 0$ into Equation (2.3.6), i.e.,

$$S_X(\omega) = A^2 |\omega|^{1-2H}, \qquad \omega \approx 0. \tag{4.2.2}$$

Hence, the spectrum of the FD process and the unaliased spectrum of FGN are identical in shape at low frequencies. However, at higher frequencies, there will be departures between the power spectral density function of FGN and the FD process (see Figure 2.6.1), and these departures will increase as H moves away from 0.5 in either direction and the dynamic range of the spectrum of both processes increases. Therefore, model misspecification appears to impact long-memory parameter estimation (either H or d) more significantly as H moves away from 0.5 (or equivalently d from 0), but is more robust otherwise.

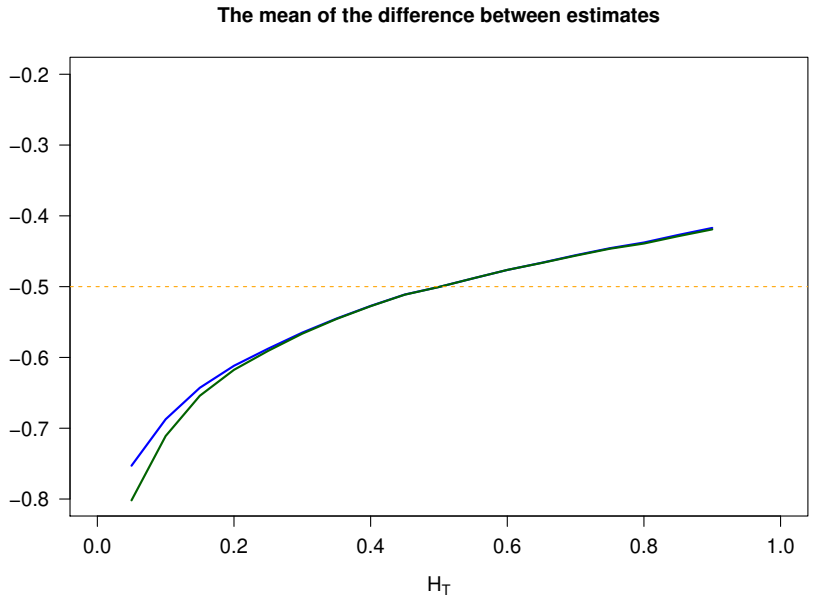

Figure 4.2.8: The line plots of the mean of $\hat{d} - \hat{H}$ calculated from each set of simulated FGNs with a fixed value of H_T varied between zero and one.

4.3 Estimation Results from Simulated FBMs

FBM is controlled by two parameters, H and A. The Hurst exponent or H describes the fractional behaviour of this process, while the parameter A specifically sets its level. By using Equations (2.2.4) and (2.3.3), this level variable can be written as a function of both the Hurst exponent (H) and the variance of its first-order difference process, denoted as σ_X^2, and this is given by

$$A = f(H, \sigma_X^2) = \frac{\sigma_X^2 \pi \Gamma(2H+1)}{\Gamma(H)\Gamma(1-H)}. \tag{4.3.1}$$

Figure 4.3.1 illustrates a line plot of A as a function of H when H is varied between zero and one, and A is in the form of the multiple of σ_X^2. When $H = 0.5$, A is exactly equal to σ_X^2, and this can also be derived from Equation (4.3.1). The relationship

between A and H is represented by a left-skewed bell-shaped curve.

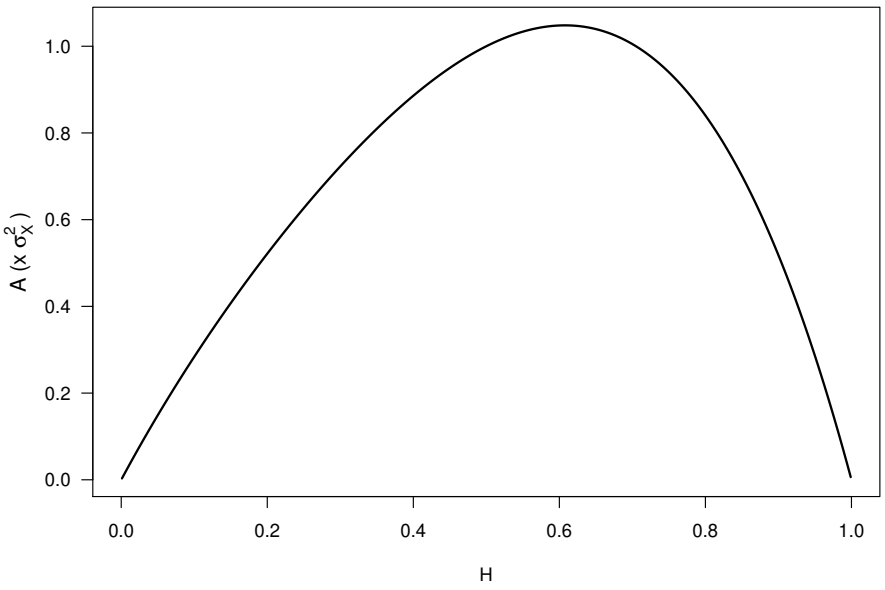

Figure 4.3.1: The level variable (A), which is in a form of the multiple of σ_X^2, as a function of the Hurst exponent (H).

We fit time series with length $n = 1000$ to each discretely-sampled FBM derived from the cumulative sum of simulated FGN used in Section 4.2. We simultaneously estimate both H and A with all five estimation methods listed in that section. For each of the five estimation methods, we consider three varying ways of estimating parameters. We shall call these three variants (a), (b), and (c), respectively, where

(a) taking a first-order difference of the data and then fitting FGN as performed in Section 4.2.

(b) fitting the data directly to FBM using the spectral density and expected periodogram from Equations (3.6.1) and (3.6.2), respectively (see Section 3.6).

(c) as in (b), but applying a 20% cosine taper as discussed in Section 3.7.

Figure 4.3.2 shows line plots of the comparison between H_T and the mean of estimated values of H using all five estimators proposed in Chapter 3. Each line represents the three varying ways of estimating parameters (a–c), as previously defined. The bias of the estimated Hurst exponent from each method and variant applied to simulated fractional processes with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ is summarised in Table 4.3.1. When $H_T > 0.5$, the periodogram without tapering in variant (b) greatly underestimates the Hurst exponent. However, the 20% cosine taper in variant (c) can reduce this bias from variant (b). According to this table, the least bias of estimated Hurst exponent from discretely-sampled FBM uses either variant (a) or (c) combined with Method 4 (the DWLE).

Table 4.3.2 presents the average computational time for parameter estimation. Each estimator, except Method 4, has similar values of computational time among all variants. However, when H_T is close to one, both variants (a) and (c) from the first three methods have high computational time, likely due to the overestimation problem of σ_X^2 for FGN (as discussed earlier), where this parameter is also required to estimate the level parameter A for FBM. Variant (b) has less computation time than variant (c) when H_T is close to one, but it provides much higher bias, as described earlier.

Method 4 is the only estimation method with much higher computational time for variants (b) and (c) than variant (a). This is due to the double summation for the nonstationary autocovariance to calculate the expected periodogram shown in

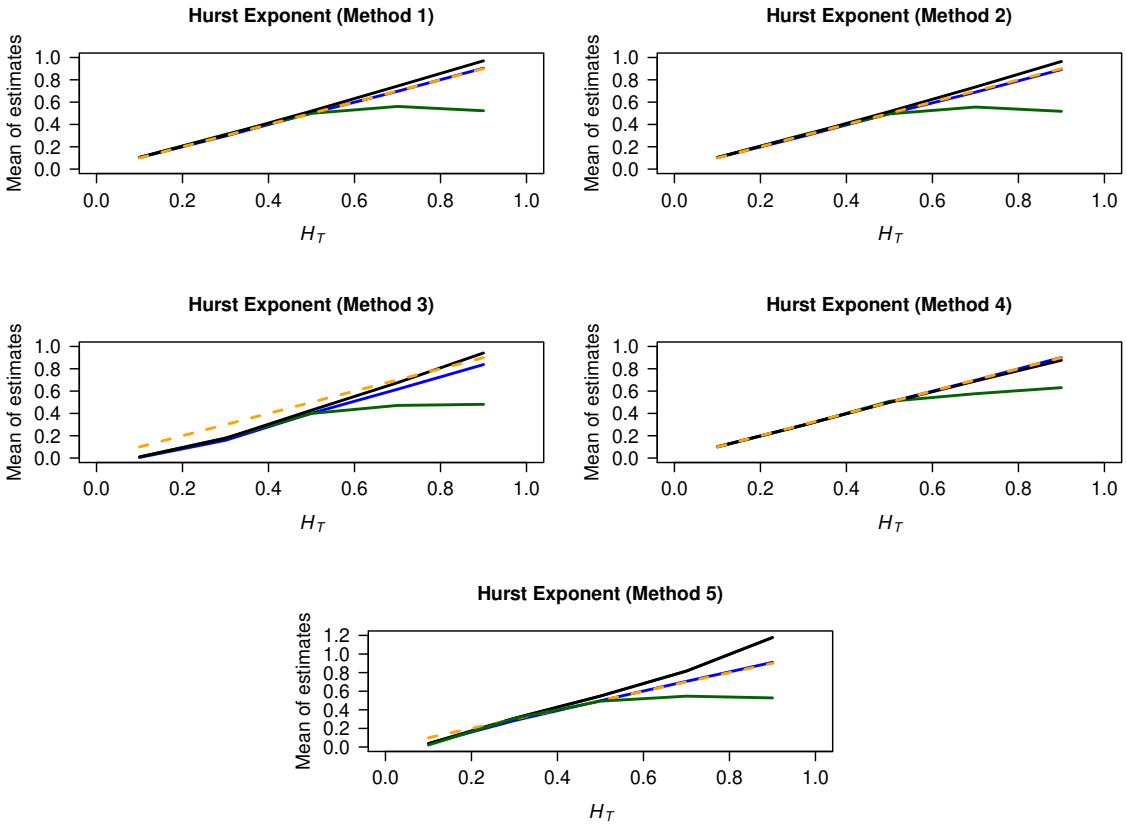

Figure 4.3.2: The line plots of the comparison between the true values of the Hurst exponent (H_T) and the mean of its estimates. All five methods are the same order as those given in Section 4.2. Each figure includes three solid lines representing three variants: (a)-blue, (b)-green, (c)-black. The orange dashed line is used as a reference line with the mean of estimates $\hat{H} = H_T$.

Equation (3.6.2). This double summation means that computation of the expected periodogram is an $\mathcal{O}(n^2 \log n)$ operation for variants (b) and (c). This is much more expensive than variant (a) using Equation (3.3.4), which is $\mathcal{O}(n \log n)$. By comparing bias and average computational time from all methods and variants, we recommend estimating the Hurst exponent by using the DWLE to fit FGN to the first-order

Table 4.3.1: Bias of estimated Hurst exponent of discretely sampled FBM.

Method	Variant	$H_T = 0.1$	$H_T = 0.3$	$H_T = 0.5$	$H_T = 0.7$	$H_T = 0.9$
1	(a)	0.002	−0.001	−0.001	−0.001	0.001
	(b)	0.005	0.011	−0.001	−0.139	−0.377
	(c)	0.005	0.006	0.020	0.044	0.070
2	(a)	0.002	−0.003	−0.006	−0.009	−0.012
	(b)	0.005	0.009	−0.005	−0.144	−0.382
	(c)	0.005	0.004	0.015	0.035	0.065
3	(a)	−0.095	−0.137	−0.001	−0.081	−0.067
	(b)	−0.097	−0.120	−0.100	−0.228	−0.419
	(c)	−0.097	−0.123	−0.071	−0.025	0.042
4	(a)	−0.001	−0.001	−0.001	−0.001	−0.001
	(b)	0.001	−0.006	0.008	−0.123	−0.269
	(c)	0.001	−0.004	−0.002	−0.010	−0.024
5	(a)	−0.073	−0.012	−0.001	0.005	0.010
	(b)	−0.078	0.008	−0.007	−0.153	−0.372
	(c)	−0.067	0.012	0.051	0.119	0.283

Table 4.3.2: Average computational time in second(s) for parameter estimation of discretely sampled FBM.

Method	Variant	$H_T = 0.1$	$H_T = 0.3$	$H_T = 0.5$	$H_T = 0.7$	$H_T = 0.9$
1	(a)	2.835	2.369	1.893	2.414	4.861
	(b)	2.926	3.111	2.275	2.525	1.986
	(c)	3.285	2.781	2.018	2.726	17.889
2	(a)	0.157	0.129	0.107	0.131	0.233
	(b)	0.155	0.145	0.118	0.136	0.110
	(c)	0.179	0.144	0.109	0.140	0.807
3	(a)	0.139	0.065	0.032	0.029	0.041
	(b)	0.447	0.064	0.037	0.041	0.028
	(c)	0.503	0.067	0.029	0.028	0.131
4	(a)	0.076	0.060	0.050	0.064	0.113
	(b)	15.733	17.385	25.534	77.682	219.638
	(c)	21.377	18.110	13.754	15.881	28.213
5	(a)	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	(b)	0.001	0.002	0.003	0.001	0.001
	(c)	0.002	0.001	0.001	0.001	0.001

difference of the data as this provides less bias and average computational time than the others.

4.4 Uncertainty Estimation

In this section, we discuss about uncertainty estimation of estimated parameters from a single simulated time series of a stationary fractional process. We assume that this process is FGN or $\{X_t\}$ with constant mean and variance over time. We consider the case where the debiased Whittle likelihood was used to estimate both H and σ_X^2 of this process. Let $\hat{\boldsymbol{\theta}} = \{\hat{H}, \hat{\sigma_X^2}\}$ be a set of estimates from this method. The standard error of these estimates can be approximated from the square root of their variances from the diagonal elements of the inverse of Fisher information matrix, i.e.,

$$\text{se}[\hat{\boldsymbol{\theta}}] \approx \sqrt{\text{Var}[\hat{\boldsymbol{\theta}}]} = \sqrt{\text{diag}(\boldsymbol{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}}))}, \tag{4.4.1}$$

where $\boldsymbol{\mathcal{F}}$ is the Fisher information matrix. This approximation comes from asymptotic theory of maximum likelihood estimators. The Fisher information matrix is defined as the negative of the expected value of the Hessian matrix such that

$$\boldsymbol{\mathcal{F}}(\hat{\boldsymbol{\theta}}) = -\mathbb{E}\left[\boldsymbol{\mathcal{H}}(\hat{\boldsymbol{\theta}})\right], \tag{4.4.2}$$

where $\boldsymbol{\mathcal{H}}$ is the Hessian matrix. Each entry is the second partial derivative of the debiased Whittle log-likelihood function, i.e., $\boldsymbol{\mathcal{H}}_{ij}(\hat{\boldsymbol{\theta}}) = \partial^2 \ell_{\text{DWLE}}(\boldsymbol{\theta}) / \partial\theta_i \partial\theta_j |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. To show the mathematical expression for each entry of the Hessian matrix, we start our process by taking the first and second partial derivatives of the autocovariance

sequence of FGN, and these are given by

$$\frac{\partial s_{X,\tau}}{\partial H} = \frac{\sigma_X^2}{2}\bigg(2|\tau+1|^{2H}\log|\tau+1| - 4|\tau|^{2H}\log|\tau| + 2|\tau-1|^{2H}\log|\tau-1|\bigg), \quad (4.4.3)$$

$$\frac{\partial s_{X,\tau}}{\partial \sigma_X^2} = \frac{1}{2}\big(|\tau+1|^{2H} - 2|\tau|^{2H} + |\tau-1|^{2H}\big), \quad (4.4.4)$$

$$\frac{\partial^2 s_{X,\tau}}{\partial H^2} = \frac{\sigma_X^2}{2}\bigg(4|\tau+1|^{2H}(\log|\tau+1|)^2 - 8|\tau|^{2H}(\log|\tau|)^2 + 4|\tau-1|^{2H}(\log|\tau-1|)^2\bigg), \quad (4.4.5)$$

$$\frac{\partial^2 s_{X,\tau}}{\partial (\sigma_X^2)^2} = 0, \quad (4.4.6)$$

$$\frac{\partial^2 s_{X,\tau}}{\partial H \partial \sigma_X^2} = \frac{1}{2}\bigg(2|\tau+1|^{2H}\log|\tau+1| - 4|\tau|^{2H}\log|\tau| + 2|\tau-1|^{2H}\log|\tau-1|\bigg), \quad (4.4.7)$$

where $|\tau| \neq 0, 1$. At $\tau = 0$, the autocovariance sequence becomes the variance of the process or σ_X^2, and all its first and second partial derivatives are zero, except that the first partial derivative with respect to σ_X^2 is one. At $|\tau| = 1$, the autocovariance sequence is given by $s_{X,\tau=1} = \sigma_X^2(2^{2H-1} - 1)$. Its first and second partial derivatives with respect to H and H^2 are $\sigma_X^2(2^{2H}\log 2)$ and $\sigma_X^2(2^{2H+1}(\log 2)^2)$, respectively, while the second partial derivative with respect to both H and σ_X^2 is $2^{2H}\log 2$. The first and second partial derivatives with respect to σ_X^2 and $(\sigma_X^2)^2$ can be directly derived from Equations (4.4.4) and (4.4.6).

Then we calculate the first and second partial derivatives of the expected periodogram as

$$\frac{\partial \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_i} = 2 \cdot \mathrm{Re}\bigg(\sum_{\tau=0}^{n-1}\bigg(1 - \frac{\tau}{n}\bigg)\frac{\partial s_{X,\tau}}{\partial \theta_i}e^{-i\omega\tau} - \frac{\partial s_{X,0}}{\partial \theta_i}\bigg), \quad (4.4.8)$$

$$\frac{\partial^2 \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = 2 \cdot \mathrm{Re}\bigg(\sum_{\tau=0}^{n-1}\bigg(1 - \frac{\tau}{n}\bigg)\frac{\partial^2 s_{X,\tau}}{\partial \theta_i \partial \theta_j}e^{-i\omega\tau} - \frac{\partial^2 s_{X,0}}{\partial \theta_i \partial \theta_j}\bigg), \quad (4.4.9)$$

where θ_i and θ_j are elements in $\boldsymbol{\theta} = \{H, \sigma_X^2\}$. These two equations are used for the computation of the first and second partial derivatives of the debiased Whittle log-likelihood function given by

$$\begin{aligned}\frac{\partial \ell_{\mathrm{DWLE}}(\boldsymbol{\theta})}{\partial \theta_i} &= -\frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}} \frac{\partial}{\partial \theta_i}\left\{\left(\log \bar{S}(\omega|\boldsymbol{\theta}) + \frac{I(\omega)}{\bar{S}(\omega|\boldsymbol{\theta})}\right)\right\} \\ &= -\frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}} \left\{\left(\frac{1}{\bar{S}(\omega|\boldsymbol{\theta})} - \frac{I(\omega)}{\bar{S}^2(\omega|\boldsymbol{\theta})}\right)\frac{\partial \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_i}\right\}, \end{aligned} \tag{4.4.10}$$

and

$$\begin{aligned}\frac{\partial^2 \ell_{\mathrm{DWLE}}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}} \Bigg\{&\left(\frac{1}{\bar{S}(\omega|\boldsymbol{\theta})} - \frac{I(\omega)}{\bar{S}^2(\omega|\boldsymbol{\theta})}\right)\frac{\partial^2 \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \\ &- \left(\frac{1}{\bar{S}^2(\omega|\boldsymbol{\theta})} - \frac{2I(\omega)}{\bar{S}^3(\omega|\boldsymbol{\theta})}\right)\left(\frac{\partial \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_i}\cdot\frac{\partial \bar{S}(\omega|\boldsymbol{\theta})}{\partial \theta_j}\right)\Bigg\}. \end{aligned} \tag{4.4.11}$$

We use Equation (4.4.11) to define the Hessian matrix $\boldsymbol{\mathcal{H}}$ for the debiased Whittle likelihood estimation.

Because the Fisher information matrix $\boldsymbol{\mathcal{F}}$ is the negative of the expected value of the Hessian matrix, each entry of $\boldsymbol{\mathcal{F}}$ with the i^{th} row and the j^{th} column given the set of estimates, $\hat{\boldsymbol{\theta}} = \{\hat{H}, \hat{\sigma_X^2}\}$, is calculated as

$$\begin{aligned}\boldsymbol{\mathcal{F}}_{ij} &= \frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}} \Bigg\{\left(\frac{1}{\bar{S}(\omega|\hat{\boldsymbol{\theta}})} - \frac{\mathbb{E}\left[I(\omega)\right]}{\bar{S}^2(\omega|\hat{\boldsymbol{\theta}})}\right)\frac{\partial^2 \bar{S}(\omega|\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i \partial \hat{\theta}_j} \\ &\qquad - \left(\frac{1}{\bar{S}^2(\omega|\hat{\boldsymbol{\theta}})} - \frac{2\mathbb{E}\left[I(\omega)\right]}{\bar{S}^3(\omega|\hat{\boldsymbol{\theta}})}\right)\left(\frac{\partial \bar{S}(\omega|\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i}\cdot\frac{\partial \bar{S}_n(\omega|\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_j}\right)\Bigg\} \\ &= \frac{1}{2}\sum_{\omega\in\boldsymbol{\Omega}} \left\{\frac{1}{\bar{S}^2(\omega|\hat{\boldsymbol{\theta}})}\cdot\frac{\partial \bar{S}(\omega|\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i}\cdot\frac{\partial \bar{S}(\omega|\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_j}\right\}, \end{aligned} \tag{4.4.12}$$

where $\hat{\theta}_i$ and $\hat{\theta}_j$ are elements in $\hat{\boldsymbol{\theta}}$. The simplification of the expression occurs because $\mathbb{E}\left[I(\omega)\right] = \bar{S}(\omega|\hat{\boldsymbol{\theta}})$. Finally, the standard error of estimates is approximately the square root of each diagonal element of the inverse of this Fisher information matrix ($i = j$), as shown in Equation (4.4.1), and the covariance between the parameter estimates can be approximated by looking at the off-diagonal entries $i \neq j$.

Figure 4.4.1 shows the line plots of standard errors of $\hat{H}$ and $\hat{\sigma}_X^2$ with respect to $\hat{H}$ from a single simulated time series of FGN. These are generated by the following order: (1) assuming that $\hat{\sigma}_X^2 = 1$ and $\hat{H}$ is varied between zero and one, (2) substituting these estimated values into Equation (4.4.12) to find the Fisher information matrix, and (3) using this matrix to calculate the standard errors of both estimates as given in Equation (4.4.1). It is reasonable that the standard error of $\hat{H}$ is low near its lower and upper boundaries due to our constrained optimisation, while this measure is maximised at $\hat{H} \approx 0.8$. The standard error of $\hat{\sigma}_X^2$ is low and constant at $\hat{H} \leq 0.7$, while it greatly jumps to a value close to $\hat{\sigma}_X^2$ when $\hat{H}$ approaches one. Thus, the DWLE may not perform well on the estimation of the sample variance of a simulated FGN with very high degree of long-range dependence.

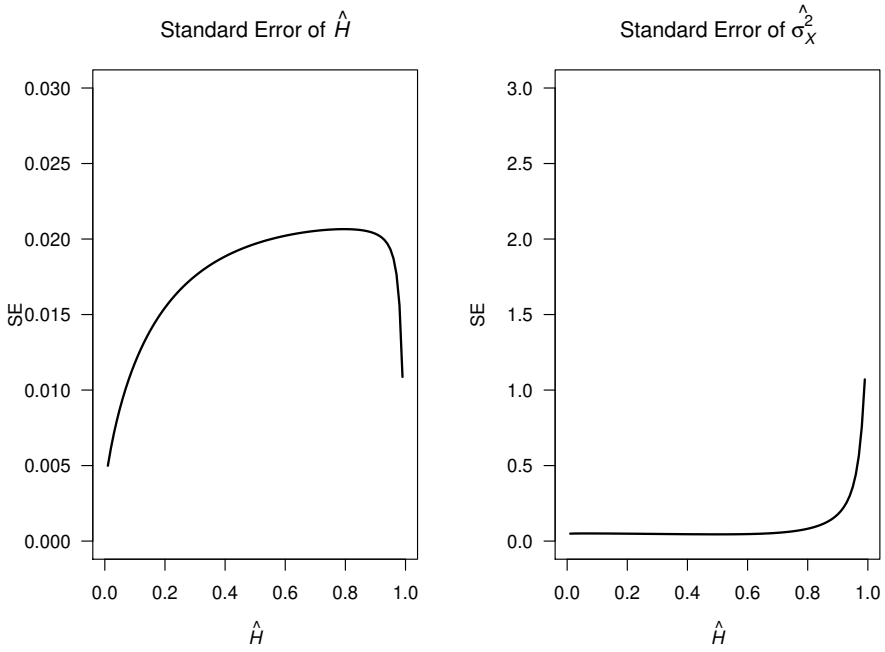

Figure 4.4.1: The line plots of standard errors of estimates from a single simulated time series of FGN.

A theoretical 95% confidence interval of estimates of both the Hurst exponent and the sample variance from simulated FGNs with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\sigma_T^2 \in \{0.25, 1, 2\}$ can be drawn by an ellipse as shown in Figure 4.4.2. Parameters from a set of 1000 time series of simulated FGNs having the same H_T and σ_T^2 were estimated with the DWLE. We compare the joint distribution of these 1000 pairs of estimates with their 95% confidence ellipse, and we find that most joint estimates from each set are inside these ellipses. There is a slight skewness of the distribution of parameter estimates from the set with $H_T = 0.9$, and this is due to the skewness of the sample variance estimates from FGNs with very high H_T, as previously discussed in Section 4.2.

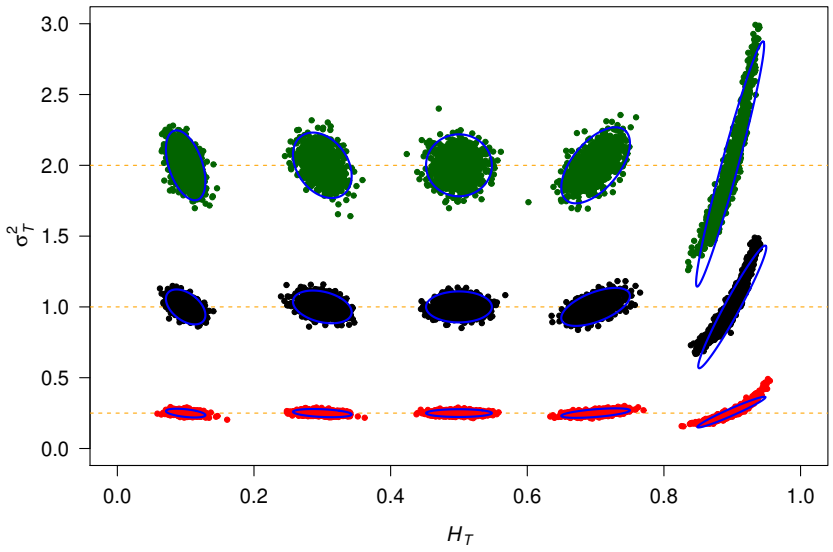

Figure 4.4.2: Points of estimates and their theoretical 95% confidence interval presented by an ellipse, from each of 1000 simulated time series of FGNs with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\sigma_T^2 \in \{0.25, 1, 2\}$.

Each confidence ellipse is constructed from the inverse of the Fisher information matrix, which is equivalent to the covariance matrix of estimates such that

$$\boldsymbol{C}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{F}}^{-1}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathrm{Var}[\hat{H}] & \mathrm{Cov}[\hat{H}, \hat{\sigma_X^2}] \\ \mathrm{Cov}[\hat{H}, \hat{\sigma_X^2}] & \mathrm{Var}[\hat{\sigma_X^2}] \end{pmatrix}. \tag{4.4.13}$$

The lengths of both major and minor axes of this ellipse are given by

$$l_{\mathrm{major}} = 2\sqrt{\chi^2_{2,\alpha}\lambda_1}, \tag{4.4.14}$$

and

$$l_{\mathrm{minor}} = 2\sqrt{\chi^2_{2,\alpha}\lambda_2}, \tag{4.4.15}$$

where λ_1 and λ_2 are the first and second largest eigenvalues of $\boldsymbol{C}(\hat{\boldsymbol{\theta}})$, and $\chi^2_{2,\alpha}$ is the critical value from the chi-square distribution with two degrees of freedom and a significance level of α. For a 95% confidence ellipse, the value of α is $1 - 0.95 = 0.05$, and thus we can substitute $\chi^2_{2,0.05} \approx 5.99$ into Equations (4.4.14) and (4.4.15). Eigenvalues λ_1 and λ_2 are given by

$$\lambda_1 = \frac{C_{11} + C_{22} + \sqrt{(C_{11} - C_{22})^2 + 4(C_{12})^2}}{2}, \tag{4.4.16}$$

and

$$\lambda_2 = \frac{C_{11} + C_{22} - \sqrt{(C_{11} - C_{22})^2 + 4(C_{12})^2}}{2}, \tag{4.4.17}$$

where C_{ij} is an element in the i^{th} row and the j^{th} column of matrix $\boldsymbol{C}$ in Equation (4.4.13). It is of note that these equations of eigenvalues are valid only when $C_{12} \neq 0$, which will be the case in practice.

The angle of the major axis of the ellipse in Figure 4.4.2 with respect to the

positive x-axis, denoted by $\hat{\psi}$, is calculated as

$$\hat{\psi} = \arctan\left(\frac{\lambda_1 - C_{11}}{C_{12}}\right), \tag{4.4.18}$$

where $\boldsymbol{v} = (C_{12}, \lambda_1 - C_{11})$ is the eigenvector corresponding to the largest eigenvalue or λ_1. According to Equation (4.4.16), it is obvious that $\lambda_1 - C_{11}$ is always non-negative. If $C_{12} = \text{Cov}[\hat{H}, \hat{\sigma_X^2}] > 0$, $\hat{\psi}$ is between 0 and $\pi/2$ radians, and this result is shown when $H_T > 0.5$. If $C_{12} = \text{Cov}[\hat{H}, \hat{\sigma_X^2}] < 0$, $\hat{\psi}$ is between $-\pi/2$ and 0 radians, and this result is shown when $H_T < 0.5$. For Gaussian white noise with $H_T = 0.5$, the theoretical value of C_{12} is 0. All off-diagonal elements of matrix $\boldsymbol{C}$ are 0, and thus its eigenvector is either $\boldsymbol{v} = (1, 0)$ or $\boldsymbol{v} = (0, 1)$ depending on whether $\text{Var}[\hat{H}]$ is greater than $\text{Var}[\hat{\sigma_X^2}]$. $\hat{\psi}$ is derived from the arctangent of the ratio of the magnitude of the y-component to the magnitude of the x-component of $\boldsymbol{v}$, so $\hat{\psi}$ is either 0 or $\pi/2$ radians for Gaussian white noise. In Figure 4.4.3, the 95% confidence interval of the correlation coefficient between $\hat{H}$ and $\hat{\sigma_X^2}$, from repeated simulation of FGNs with $\sigma_T^2 = 1$ and H_T varied between zero and one, is shown by the shaded region, while the correlation coefficient derived from the covariance matrix is drawn by the curved black line. The sign of the correlation coefficient from this line corresponds to that of the angle $\hat{\psi}$ of ellipse with the same value of H_T shown in Figure 4.4.2.

The length of projection vector of the major axis of the ellipse on either axis of parameter is directly proportional to the standard error of estimates of such parameter. For each column in Figure 4.4.2 with the same H_T, the projection vectors of the major axes of all ellipses on the horizontal axis are equal in length. However, the lengths of their projection vectors on the vertical axis are different but scaled by

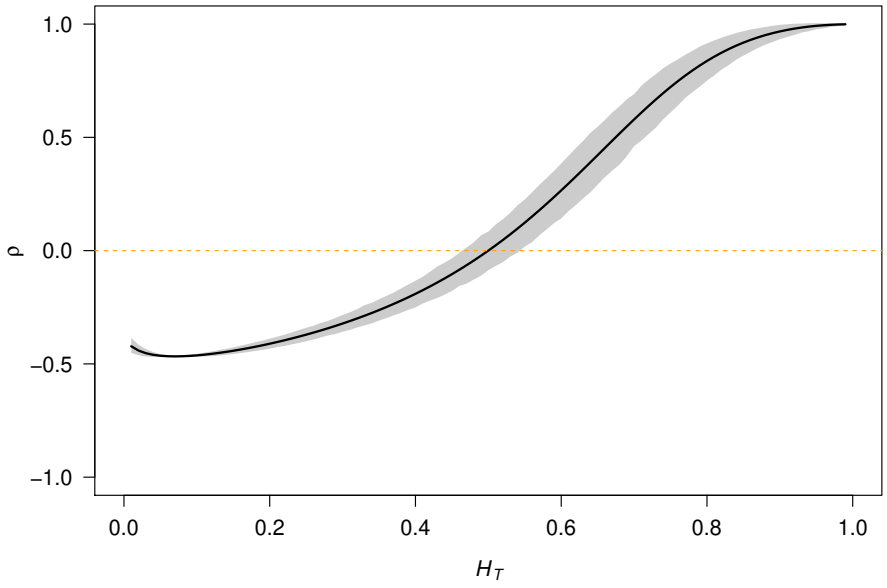

Figure 4.4.3: A line plot and the shaded region representing the correlation coefficient between both parameters along with its 95% confidence interval using the DWLE.

σ_T^2. Let us compare two different FGNs. The first process provides a set of estimates as $\{\hat{H}, \hat{\sigma_X^2}\}$ from a set of their true values or $\{H_T, \sigma_T^2\}$. The second process has the same H_T but its true value of the sample variance is multiplied by k, i.e., $k\sigma_T^2$. The expected periodogram and the first partial derivative of the autocovariance sequence with respect to the Hurst exponent of the second process is also increased by a factor of k. By using Equations (4.4.8), (4.4.9), (4.4.12), and (4.4.13), We calculate the covariance matrix of this process, and it is given by

$$\boldsymbol{C}^{(k)}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathrm{Var}[\hat{H}] & k \cdot \mathrm{Cov}[\hat{H}, \hat{\sigma_X^2}] \\ k \cdot \mathrm{Cov}[\hat{H}, \hat{\sigma_X^2}] & k^2 \cdot \mathrm{Var}[\hat{\sigma}_X^2] \end{pmatrix}. \tag{4.4.19}$$

Hence, the standard error of estimated values of the Hurst exponent is unchanged, while the standard error of estimated values of the sample variance is multiplied by

k. These are visualised by the lengths of projection vectors, as previously discussed.

Overall, this section has demonstrated that the asymptotic approximation of parameter estimate uncertainty via the Fisher Information matrix provides a good mechanism for reporting parameter uncertainties from the DWLE applied to FGN.

4.5 Application on Log-Volatility Time Series

In this section, we applied the DWLE along with its uncertainty estimation to the log-volatility time series of financial assets. Gatheral et al. (2018) evaluated the smoothness of these time series from several financial assets. They found that these time series behave like fractional Brownian motion with the Hurst exponent between 0.08 and 0.2. This means that the log-volatility time series are likely to exhibit roughness as compared with standard Brownian motion. For this reason, they proposed a new model called the rough fractional stochastic volatility model to fit these time series. However, in this section we modelled these same time series as exact FBM processes by fitting FGN to the first-order difference time series and estimated the *local* Hurst exponent at each time point of the financial data. By "local" we mean a time-varying Hurst exponent which is calculated over rolling windows of fixed width. The data we shall consider here are the log-volatility time series of four stock indices including S&P 500, FTSE 100, DAX 30, and NIKKEI 225 from the first trading day of 2011 to the last trading day of 2020, and they are all reported in the Oxford-Man Institute's realized library (https://realized.oxford-man.ox.ac.uk). These time series are presented in Figure 4.5.1. To estimate the local Hurst exponent at each trading

day, we use a moving window to filter the time series with the DWLE. Specifically, we use a moving window with length $n = 256$, which is also the approximate number of trading days per year. This local Hurst exponent is estimated from the window of the first-order difference of log-volatility time series from the previous 255 days until the day of interest. In addition, we estimated the global Hurst exponent from these 10-year time series of financial assets.

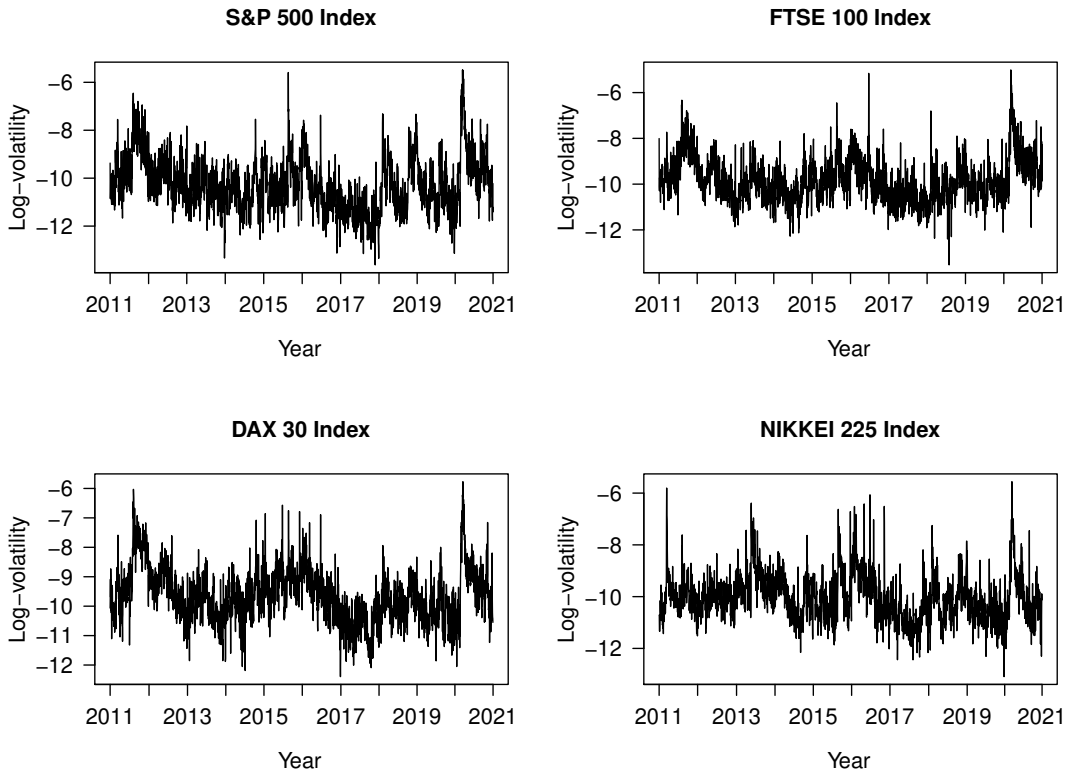

Figure 4.5.1: Log-volatility time series of S&P 500, FTSE 100, DAX 30, and NIKKEI 225 from the first trading day of 2011 to the last trading day of 2020.

The estimated results of the local Hurst exponent and its corresponding 95% confidence intervals are illustrated in Figure 4.5.2. We also provide the estimate of the global Hurst exponent for each stock index. Although our observed period of

time series is different from that in Gatheral et al. (2018), the global Hurst exponent is still about 0.1, while the local Hurst exponent fluctuates around this value of the global parameter. The global Hurst exponent from log-volatility time series of S&P 500 is 0.1644, which is slightly higher than that from others in our study and the same financial indices in different periods from Gatheral et al. (2018) ($H = 0.1420$). We also observe that the drastic change of the pattern of the local Hurst exponent of each plot in this figure is related to the change of the stock price due to some major events. We list two examples of these as follows: The stock market crash due to the COVID-19 pandemic increases the fluctuation level of the local Hurst exponent near the beginning of 2020 for all stock indices. The uncertainty of investors about the Brexit deal in the beginning of 2018 causes a sudden increase including a single peak of the local Hurst exponent for the FTSE 100 index.

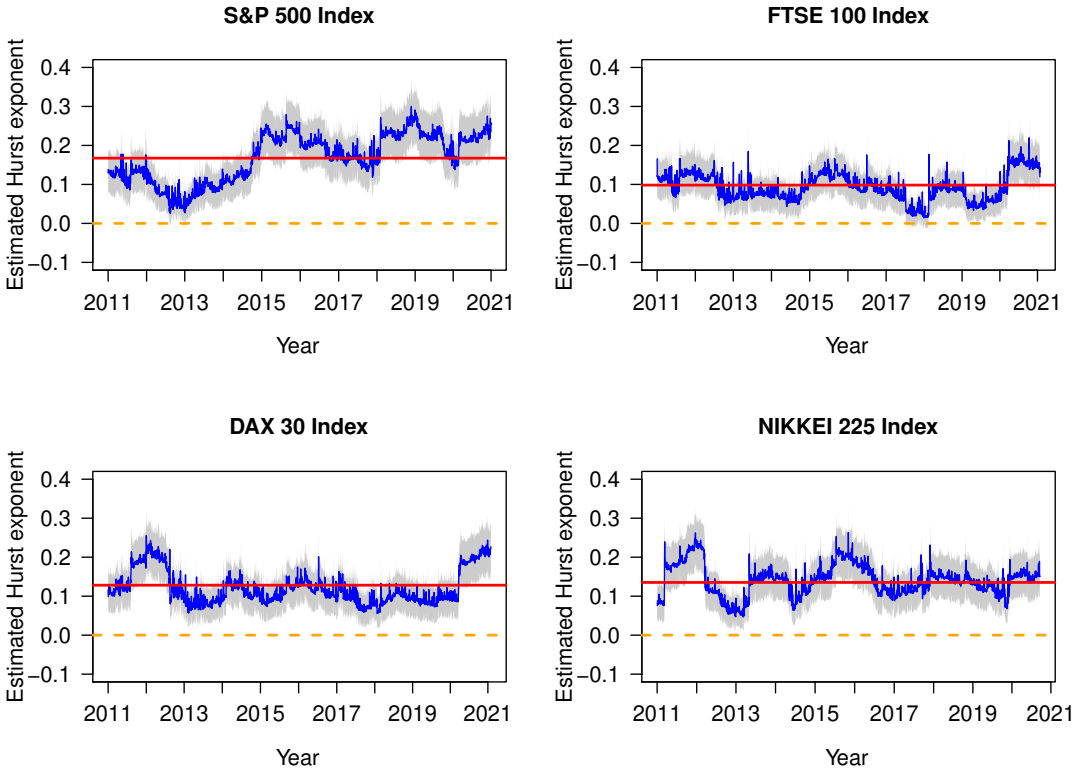

Figure 4.5.2: The estimation of the local Hurst exponent and its 95% confidence interval at each working day of four stock indices from 2011 to 2020 (blue solid line and grey shaded region), the estimation of the global Hurst exponent from this whole range of the log-volatility time series (red solid line), and the reference line of $H = 0$ (orange dashed line).

Chapter 5

Nonparametric Statistics for High-Frequency Accelerometry Data from Individuals with Advanced Dementia

Accelerometry data provides a method for monitoring physical activity over time. Such data is typically collected from an actigraph device, which is generally worn on the wrist, for a continuous period of time with minor impact on daily life. The device contains an accelerometer which measures acceleration of the individual, and sometimes a light and temperature sensor is also included. Data recorded from such instruments has been widely used to study the pattern of 24-hour rest-activity rhythm, also known as circadian rhythm, of humans for almost half a century (Ancoli-Israel

et al., 2003). Of primary interest here is the increased use of actigraphy and accelerometry with people with advanced dementia at the end of life (Kok et al., 2017; Khan et al., 2018). However, the analysis of such data from this population where circadian rhythms are significantly dysregulated is challenging, putting into question the validity of conclusions extracted.

Several measures have been proposed to quantify the circadian rhythm and they can be broadly classified into two groups: parametric and nonparametric measures. Cosinor is a traditional parametric method for calculating the amplitude and phase of the circadian rhythm (Halberg et al., 1967). This procedure fits the observed data by a regression model of continuous cosine functions with a rhythm-adjusted mean, where the period is assumed to be known (Cornerllissen, 2014). This model is often a poor fit to accelerometry data, especially for individuals with weak circadian patterns (Fossion et al., 2017). Alternative methods include singular spectrum analysis (SSA) which employs periodic components with varied amplitude and phase to fit the data (Fossion et al., 2017), or the use of multiparameter-extended cosine functions (Krafty et al., 2019). In general, these parametric methods can be very useful for characterising circadian rhythms, however they are typically model-based and suffer the usual disadvantages of parametric methods: namely the bias and lack of robustness that results when model assumptions are not met by the application data source of interest.

Nonparametric measures are model-free and require little to no user expertise to implement. Two such important and widely-used measures are interdaily stability (IS) and intradaily variability (IV). These were first proposed in 1990 to study how

the circadian rhythm changes for patients with aging and Alzheimer's disease (Witting et al., 1990). IS measures the strength of the circadian rhythm, while IV measures the *fragmentation* of the data by measuring the variation of hour-to-hour data relative to sample variance. In Witting et al. (1990) and other such early studies, both IS and IV were calculated from data with an hourly sampling interval due to the limitations of sensor processing in the actigraph device. However, storage capacity has now been increased and data can be recorded with much higher temporal sampling intervals (Gonçalves et al., 2014). This is important as IV values change significantly as a function of the temporal sampling rate, although we note that the rate only has a very small effect on IS in general. In this chapter, we study the effect of temporal sampling on IV in detail, and propose an optimal rate, together with a simple method that ensures no data is thrown away in calculating the statistic.

IS and IV have been used as summary statistics for actigraphy and accelerometry data from individuals with dementia in several studies. A number of studies have reported that patients with dementia have significantly lower IS than those without dementia (Witting et al., 1990; Harper et al., 2001; Satlin et al., 1995; Hatfield et al., 2004), and this can be associated with cerebral microbleeds (Zuurbier et al., 2015), occipital periventricular and frontal deep white matter hyperintensities (Oosterman et al., 2008), and loss of medial temporal lobe volume (Van Someren et al., 2019). On the other hand, these patients have significantly higher IV (Witting et al., 1990; Harper et al., 2001; Hatfield et al., 2004; Hooghiemstra et al., 2015) which was positively correlated with the ratio of phosphorylated tau_{81} (pTau) to cerebrospinal fluid amyloid β 42 (Aβ 42) for the biomarker collection assessing preclinical Alzheimer dis-

ease (Musiek et al., 2018), and medial lobe atrophy (Van Someren et al., 2019). More broadly, dementia can affect the disruption of circadian rhythm which is linked with dementia biomarkers (Smagula et al., 2019).

Another nonparametric method is detrended fluctuation analysis (DFA), which was introduced to study fractal scaling behaviour and determine long-range autocorrelation in time series data (Peng et al., 1994). With DFA, a scaling exponent is estimated, and this is related to the well-known Hurst exponent H (proposed by Hurst (1951)). The Hurst exponent value is between zero and one, and this can be divided into three cases: (1) $0.5 < H < 1$ for persistent time series whose increments have positive long-term autocorrelation, (2) $0 < H < 0.5$ for anti-persistent time series whose increments have negative long-term autocorrelation, and (3) $H = 0.5$ for a Brownian process with uncorrelated "white noise" increments. The DFA scaling exponent (α), on the other hand, is defined in the range $0 < \alpha < 2$ (Ihlen, 2012). Stationary time series (such as a correlated noise) have α values between zero and one, whereas nonstationary time series (such as a random walk) have α values between one and two. The DFA scaling exponent is the same as the Hurst exponent in the stationary case such that $\alpha = H$, but in the nonstationary case we have $\alpha = H + 1$. The boundary between stationary and nonstationary time series behaves like "pink" noise (or 1/f noise) with α approximately equal to one. The DFA method has been commonly applied to time series with monofractal structure which is defined by a single scaling exponent. However, there might be some temporal fluctuations (multifractal structure) in the time series and this requires a group of generalised scaling exponents to explain them. Such fluctuations can be analysed by a method called multifractal

detrended fluctuation analysis (MFDFA) (Ihlen, 2012; Kantelhardt et al., 2002). We note that there also exist other procedures for quantifying fractal-type behaviour in a time series (see e.g. Fernández-Martínez et al. (2019)).

For its part, both DFA (or monofractal DFA) and MFDFA have been widely used to estimate scaling and Hurst exponents in many research fields. Indeed, a few research studies collected data from patients with dementia and analysed their fractal scaling behaviours using DFA (Hu et al., 2013, 2016; Huber et al., 2019). Some of their findings were that the change of DFA scaling exponent could be found from antemortem actigraphy records of some patients with dementia and its degree of change was negatively correlated with the number of two major circadian neurotransmitters found in the suprachiasmatic nucleus (Hu et al., 2013), and the timed bright light therapy reduced the rate of decay of the Hurst exponent over time (Hu et al., 2016). In this chapter, we will further explore the use of monofractal DFA on accelerometry data, including a novel comparison of considering DFA scaling exponents during daytime and nighttime separately.

Finally, spectral analysis can be used to decompose variability in accelerometry data across different frequencies of oscillation. The simplest nonparametric estimator is the periodogram (Percival and Walden, 1993). For people who do not suffer from dementia or sleep disorders, both methods usually provide that the highest point in the spectrum generally occurs around the frequency of 1/24 hour^{-1} (known as the fundamental frequency), and further peaks occur at *harmonic* frequencies, which are positive integer multiples of the fundamental frequency. Harmonic peaks occur as the circadian rhythm is not perfectly sinusoidal. For individuals with disrupted sleep

cycles, the spectrum will have relatively low energy at the fundamental and harmonic frequencies, and relatively high and noisy levels at other frequencies. In this chapter, we propose a novel nonparametric estimator from the periodogram which calculates the ratio of variability at fundamental and harmonic frequencies versus the whole spectrum. We call this method proportion of variance (PoV), and we will analyse this to our dataset and compare with other methods in order to assess the efficacy of spectral analysis methods for accelerometry data.

5.1 Materials and Methods

5.1.1 Accelerometry data

The studied dataset contains accelerometry time series from 26 individuals suffering from advanced dementia. To create a reference group without dementia, we also studied 14 recordings of accelerometry data from members of the research team and colleagues that provided this data as part of piloting the use of the actigraph device for our main study. The 26 individuals with advanced dementia took part in a group intervention project aimed at improving quality of life when living in care homes (Froggatt et al., 2020). Fifteen of them received the Namaste Care intervention in their treatment during the period of data collection, while the remaining 11 participants were allocated to a non-intervention group receiving treatment as usual (see protocol details in Froggatt et al. (2018)). Although the intervention did not have a significant effect on accelerometry readings of those with advanced dementia (Froggatt et al., 2020), we nonetheless compared intervention and non-intervention groups as a useful

indicator of the performance and stability of our time series metrics. All the participants with advanced dementia were recruited from six nursing homes between the end of 2017 and May 2018. The study end date was 30/11/2018. Being a permanent resident in a nursing care home, lack of mental capacity, and a FAST score of $6-7$ reflecting advanced dementia status were part of the inclusion criteria (see full details in Froggatt et al. (2018)). All 26 participants with advanced dementia and with valid accelerometry information are part of a larger study detailed in Froggatt et al. (2018), the demographic details of which are also included in Table A.1.1 in Appendix A for reference.

The original trial was approved by the Wales Research Ethics Committee 5 Bangor Research Ethics Committee (reference number 17/WA/0378) on 22 November 2017. Potential participants were screened by the principal investigator and the senior care team, and eligible participant's written consent was provided by a personal consultee. For those without a personal consultee, the consent was requested to a nominated consultee following care home procedures. Following this, researchers discussed the study with consultees and gained assent from residents to take part in the study.

For each participant across the study, the accelerometry data was recorded by a wrist-worn accelerometer called GENEActiv for a maximum of 28 days. This device measures acceleration in the unit of gravitational force (g-force). The sampling period of recording was set to five seconds such that the length of accelerometry time series was 17280 per day. Raw data output from each sampling time consists of three non-negative values representing the magnitudes of projected vectors of acceleration on three axes in Euclidean space. The Euclidean norm was calculated from these

magnitudes. Because the accelerometer is affected by the gravitation of Earth, this effect is removed. Thus, we calculate what is known as the *Euclidean norm minus one* (ENMO) (Van Hees et al., 2013), which results in the time series of accelerations used for implementing our nonparametric time series methods. We emphasise that in the literature other forms of actigraphy time series are sometimes studied (Ancoli-Israel et al., 2003) such as *time above threshold* (total amount of time such that the accelerometry data is above a setting threshold per time interval), *zero-crossing method* (total count of movements per time interval by counting the number of times when the accelerometry data is close to zero), and *digital integration* (total area under the curve of accelerometry data per time interval). We chose to run our analysis methods on ENMO data as this retained the most information from the raw data, but our methods could also have been applied to the types of actigraphy data described above. Before statistical analysis was performed, we excluded all data for any day having at least one missing data. This was done because the methods we present require complete measurements, and the alternative of using interpolation methods will perform poorly if there are long consecutive periods with missing data. We apply four different nonparametric time series methods (implemented in RStudio using R version 4.0.0) to the ENMO data which we now describe. R software for implementing each method, for any generic accelerometry time series, is publicly available at www.github.com/suibkitwanchai-k/Accelerometry.

5.1.2 Interdaily stability (IS)

The first statistical method is interdaily stability (IS), which is a nonparametric measure for the strength of the circadian rhythm in the accelerometry data. The circadian rhythm is driven by a circadian clock with approximately 24-hour time period. For each individual, we define $\{X_t\}$ as a time series of ENMO data with length N. The interdaily stability is the ratio of the average square-error of hourly means from the overall mean to the variance of $\{X_t\}$, i.e.,

$$\text{IS} = \frac{\sum_{s=1}^{24}(\overline{X_s} - \overline{X})^2/24}{\sum_{t=1}^{N}(X_t - \overline{X})^2/N}, \tag{5.1.1}$$

where $\overline{X_s}$ is the sample mean of ENMO data during the s^{th} hour of the day ($s = 1, 2, ..., 24$) and $\overline{X}$ is the overall sample mean. Equation (5.1.1) calculates IS by aggregating data on an hourly basis in the numerator into 24 bins. The original motivation for this is that the time series itself is sampled on an hourly interval (Witting et al., 1990) such that more bins are not possible. However for high frequency data, such as the 5-second data in our study, the hourly aggregation becomes an arbitrary choice, and IS could instead be calculated by aggregating over shorter time intervals and into several more bins. According to Gonçalves et al. (2014), however, the interdaily stability is insensitive to this choice, and we found the same in our study. Therefore we keep to the choice of hourly binning in Equation (5.1.1). In contrast, the value of intradaily variability (IV) depends strongly on the sampling period as we now discuss.

5.1.3 Intradaily variability (IV)

Intradaily variability (IV) is used to estimate circadian fragmentation in the data. In our implementation, we ensure all data in the high frequency time series is used, even if subsampling is applied. Specifically, we let

$$\{Y_k^{[j]}\} = \{X_{(k-1)\Delta+j, k=1,2,\ldots,\lfloor N/\Delta \rfloor}\} \tag{5.1.2}$$

be a finite sequence obtained from subsampling the time series of ENMO data $\{X_t\}$ with length N by an arbitrary positive integer Δ not greater than N, and j is the index of the j^{th} time series data derived from this subsampling ($1 \leq j \leq \Delta$). For each $\{Y_k^{[j]}\}$, the intradaily variability is calculated by the ratio of variation in consecutive time intervals to the total variance, i.e.,

$$\text{IV}[j] = \frac{\sum_{k=2}^{M} (Y_k^{[j]} - Y_{k-1}^{[j]})^2/(M-1)}{\sum_{k=1}^{M} (Y_k^{[j]} - \overline{Y^{[j]}})^2/M}, \tag{5.1.3}$$

where $M = \lfloor N/\Delta \rfloor$ is the length of $\{Y_k^{[j]}\}$ and $\overline{Y^{[j]}}$ is its overall mean. The final IV value is then the average of all IV values obtained from Equation (5.1.3), i.e.,

$$\text{IV} = \frac{\sum_{j=1}^{\Delta} \text{IV}[j]}{\Delta}. \tag{5.1.4}$$

This implementation, which averages across different start points of the time series, ensures all data is used regardless of the choice of Δ. Note that this procedure was also proposed in Zhang et al. (2005) for calculating integrated volatility in financial time series, which is a very similar metric to IV.

In Witting et al. (1990), the subsampling period was set to 60 minutes. However, for higher frequency time series, it can be adjusted to any other appropriate time (Fos-

sion et al., 2017; Gonçalves et al., 2014). In the next section we shall investigate in more detail the relationship between IV and the choice of Δ for our studied dataset. We shall show that there is a natural trade-off to be balanced in that overly low values for Δ lead to poor estimates due to short-lag autocorrelations, whereas large values for Δ fail to capture the differences in fragmentation for individuals with and without advanced dementia. Another consideration we make is the correlation between IV and other statistical measures, and whether this metric responds as it should to other summary statistics for a given subsampling period.

5.1.4 DFA scaling exponent

The Hurst exponent (H) ranges between zero and one, and it is used to measure the degree of long-range dependence of a time series, which is the rate of decay of its long-lag autocorrelation. A higher Hurst exponent indicates that the time series data is more persistent and has higher degree of positive long-term autocorrelation. Detrended fluctuation analysis (DFA) can be used to estimate H via the DFA scaling exponent (α) which ranges between zero and two such that $\alpha = H$ for stationary noise-like behaviour and $\alpha = H + 1$ for nonstationary random walk-like behaviour. Here we succinctly summarise the estimation of the DFA scaling exponent for our data in the following steps (see also Ihlen (2012)):

1. For a time series of ENMO data $\{X_t\}$ with length N, we construct a new time series $\{Z_t\}$ which is the cumulative sum of $\{X_t\}$ with mean centering, i.e.,

$$Z_t = \sum_{u=1}^{t} (X_u - \overline{X}), \tag{5.1.5}$$

where $1 \leq t \leq N$.

2. We divide $\{Z_t\}$ into equally-sized non-overlapping segments with length S (scale parameter) where S is chosen to vary across a range. According to Ihlen (2012), a minimum size of S of larger than 10 is considered to be a rule of thumb. For our accelerometry data, we set $S = 2^i$, where i is varied from 4 to 8 in increments of 0.25. Since our sampling period is 5 seconds, this range of S is therefore between 1.333 and 21.333 minutes. Larger values of S could have been applied to our data but we found this makes the method more computationally expensive without significantly changing the estimated exponent. A least squares regression line is then fitted to the time series data in each segment for each value of S. The local root-mean-square deviation at segment number K with length S is calculated as

$$\text{RMSD}_{S,K} = \sqrt{\frac{\sum_{j=S(K-1)+1}^{SK} (\hat{Z}_j - Z_j)^2}{S}}, \tag{5.1.6}$$

where $\hat{Z}_j$ is the predicted value from the regression line and $1 \leq K \leq \lfloor N/S \rfloor$.

3. For each S, the overall root-mean-square deviation (F) is calculated as

$$F(S) = \sqrt{\frac{\sum_{i=1}^{\lfloor N/S \rfloor} \text{RMSD}_{S,i}^2}{\lfloor N/S \rfloor}}. \tag{5.1.7}$$

4. The DFA scaling exponent (α) is estimated by the slope of the linear regression line between $\log_2(S)$ (x-axis) and $\log_2(F)$ (y-axis).

In our analysis, the DFA scaling exponent is estimated individually for each participant across the whole time series. We also calculate separate values for each

participant from daytime and nighttime readings. This is done to examine whether there is a difference between daytime and nighttime activity for any given participant, and to see whether this difference is more pronounced in any of the participant groups, thus providing more statistical insight from this type of analysis.

5.1.5 Proportion of variance (PoV)

The power spectral density (PSD) or power spectrum describes the distribution of variability in a time series as a function of frequency. It is equivalent to Fourier transform of the autocovariance sequence of a stationary time series. The periodogram is an asymptotically unbiased estimate of the PSD which is defined by

$$I(f) = \frac{1}{N}\left|\sum_{t=1}^{N}(X_t - \overline{X})e^{-i2\pi ft}\right|^2, \tag{5.1.8}$$

where N is the length of ENMO time series $\{X_t\}$, $\overline{X}$ is its overall mean, and f is the frequency in cycles per second or hertz (Hz). $i \equiv \sqrt{-1}$ represents the imaginary unit and $|\cdot|$ denotes the absolute value. For $\{X_t\}$ with real-valued data, the periodogram is real-valued and symmetric, i.e. $I(f) = I(-f)$. In addition, as $\{X_t\}$ is a discretely observed time series from sampling, then its periodogram is only defined for positive and negative frequencies that are less than half of the sampling rate, which is also called the Nyquist frequency (f_n). For our data, which was sampled every 5 seconds (sampling rate = 1/5 Hz), the Nyquist frequency is therefore equal to $f_n = 1/10$ Hz. The total area under the spectral line of the periodogram between $-f_n$ and f_n is approximately equal to the variance of $\{X_t\}$.

The periodogram can be used to assess the strength of circadian rhythm by eval-

uating its spectral value at the frequency $f = 1/24$ hour^{-1} (1/86400 Hz). However, because the total variance of ENMO data will vary across individuals, directly comparing the spectral values of the periodogram at just one frequency will be a noisy and unreliable measure of circadian strength. Instead, we propose a new method by calculating the proportion of variance (PoV) explained by the periodogram at or *near* the fundamental frequency (also called the first harmonic) of circadian rhythm ($f = 1/24$ hour^{-1}). Furthermore, because circadian patterns will typically not be perfect sinusoids, we also include variability from the three following harmonic frequencies. The PoV statistic is then obtained by finding the ratio of area under the spectral line around the frequencies of interest to the total sample variance. Specifically, we use the range of frequencies between $|f| = 1/24.5$ hour^{-1} (1/88200 Hz) and $|f| = 1/23.5$ hour^{-1} (1/84600 Hz) along with their positive integer multiples from two to four. In the simplest case of considering only the fundamental frequency and ignoring higher orders of harmonics, the PoV value for each individual is calculated as

$$\text{PoV}^{(F)} = \frac{\int_{-1/84600}^{-1/88200} I(f)df + \int_{1/88200}^{1/84600} I(f)df}{\text{Var}[X_t]} = \frac{2\int_{1/88200}^{1/84600} I(f)df}{\text{Var}[X_t]}, \tag{5.1.9}$$

where f is the frequency in hertz and $\text{Var}[X_t] = \sum_{t=1}^{N}(X_t - \overline{X})^2/(N-1)$. The superscript "$(F)$" indicates that this PoV value is calculated only near the fundamental frequency. The values from negative frequencies are included in the numerator in Equation (5.1.9) as these contribute to the total variance. Due to the symmetry of the periodogram, this contribution is equivalent to the corresponding contribution from positive frequencies which yields the simpler final expression. Therefore $\text{PoV}^{(F)}$ can be interpreted as the proportion of variance in the time series explained by abso-

lute frequencies between 1/24.5 hour^{-1} and 1/23.5 hour^{-1}. In the case of considering the first four harmonic frequencies, the PoV value for each individual is calculated as

$$\text{PoV}^{(H)} = \frac{\sum_{k=1}^{4} \int_{-k/84600}^{-k/88200} I(f)df + \sum_{k=1}^{4} \int_{k/88200}^{k/84600} I(f)df}{\text{Var}[X_t]} = \frac{2\sum_{k=1}^{4} \int_{k/88200}^{k/84600} I(f)df}{\text{Var}[X_t]}, \quad (5.1.10)$$

where again f is the frequency in hertz and the steps in the derivation follow the same pattern as Equation (5.1.9). The superscript "(H)" indicates that this PoV value is calculated near both the fundamental frequency, which is also called the first harmonic, and several higher-order harmonics. The number of harmonics used can be amended if necessary, but we found 4 to be a good value to use in practice.

There are several alternatives to using the periodogram, including smoothed and multi-tapered approaches. However, these techniques smooth across frequency leading to a loss in data resolution. Since our statistics in Equations (5.1.9) and (5.1.10) are already implicitly smoothing across frequencies near the peak frequency (and harmonics) to account for noise and variability, then we found no such further smoothing is required.

5.2 Results and Discussion

5.2.1 Accelerometry data

In the left column of Figure 5.2.1, we display three examples of time series of ENMO data, one from each of the different groups (non-intervention, intervention, and without dementia). The ENMO values from individuals with advanced dementia (non-intervention and intervention groups) were largely between 0 and 0.2g with occasional

spikes above this range, and this was consistent with other participants in the study from these groups. The data from the individual without dementia had higher ENMO values and its daily circadian pattern was more clearly observed. In the right column of Figure 5.2.1, we display the 24-hour time average plots of ENMO data from these three individuals. The average data from the individual without dementia had high ENMO values with large fluctuation during daytime and low ENMO values with moderate fluctuation during nighttime. The average data from individuals with advanced dementia, however, had low ENMO values throughout the day and no clear circadian cycle was present.

5.2.2 Interdaily stability (IS)

We used IS to evaluate the strength of the circadian rhythm in the ENMO data. The bar chart in Figure 5.2.2 displays IS values from all participants in the study. All participants without dementia except the 7^{th} and 14^{th} recordings had higher IS values than all participants with advanced dementia. The corresponding box plot in Figure 5.2.2 shows that participants in the intervention group had slightly higher mean IS value than in the non-intervention group, but this was not found to be significantly different (p-value = 0.413). However, the mean IS value from the group without dementia was significantly higher than from the combined non-intervention and intervention group (p-value < 0.001). All p-values in this chapter were calculated using the nonparametric Mann-Whitney U-test at a 5% significance level. Table A.2.1 in Appendix A presents mean, standard deviation, minimum and maximum values of all our metrics across all participant groups.

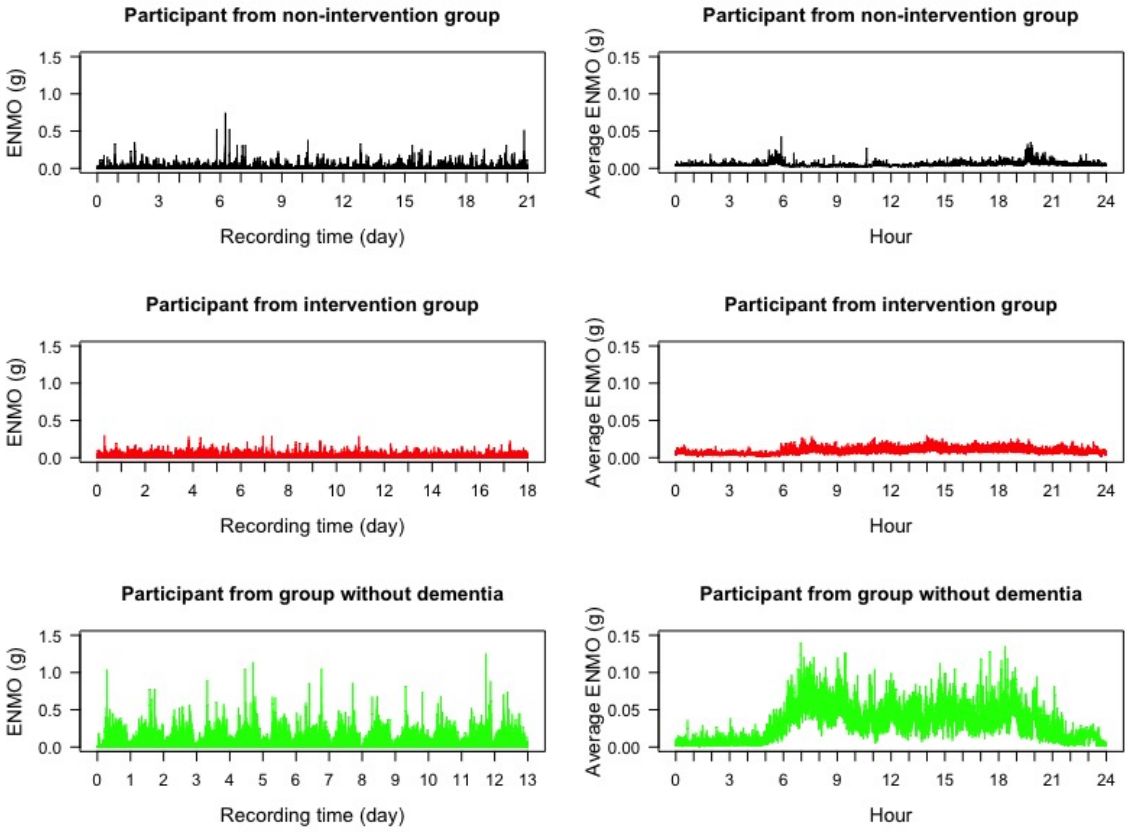

Figure 5.2.1: Time series plots of ENMO data (left) and their 24-hour time average plots (right) from three participants from the study: one from each of the non-intervention group (black), the intervention group (red), and the group of individuals without dementia (green).

5.2.3 Intradaily variability (IV)

We used IV to assess the circadian fragmentation in the ENMO data. First we attempted to find the appropriate subsampling interval with which to calculate IV in Equations (5.1.2), (5.1.3), and (5.1.4). To do this, we subsampled the ENMO data with a sampling period ranging from 5 seconds ($\Delta = 1$) to 60 minutes ($\Delta = 720$), in intervals of 5 seconds, and calculated IV with each sampling period. Due to the

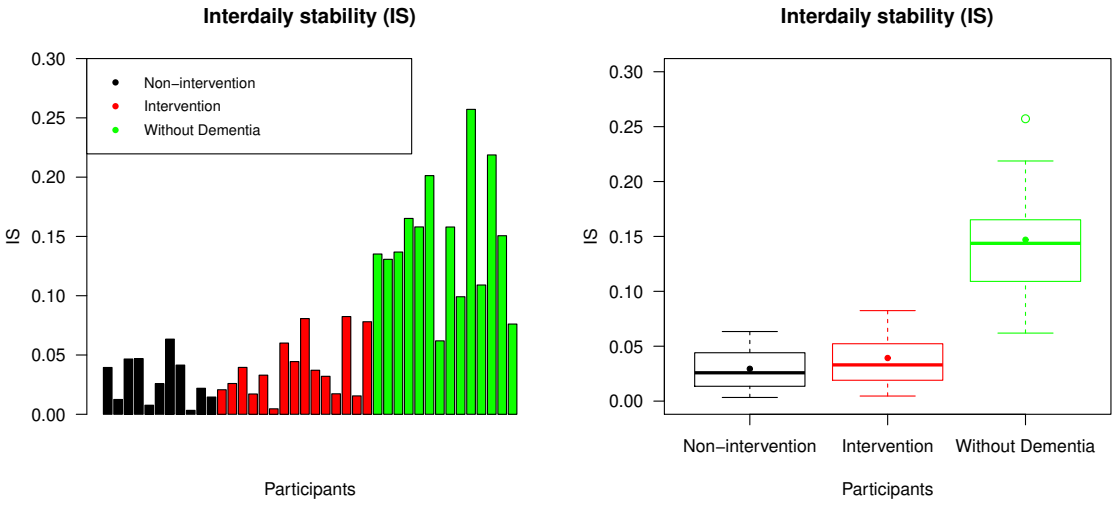

Figure 5.2.2: The left panel displays IS values across all participants and the right panel is a box plot collating these values across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group, the median and mean of the IS values are presented by a thick line and a point in its own box, respectively.

high sampling frequency of our data, the conventional subsampling interval, which is hourly subsampling, was not found to be appropriate. The left plot of Figure 5.2.3 shows the relationship between the average IV value and the subsampling interval for the three different groups of individuals. IV values increased as a function of subsampling interval for each group, converging to one another as the subsampling interval approached one hour. This pattern of IV as a function of subsampling interval is generally consistent with Figures 2, 3, and 5 of Gonçalves et al. (2014), but other more "U-shaped" relationships are found in Figure 4 of Gonçalves et al. (2014) and Figure 6 of Fossion et al. (2017). We note however that these studies were on a range of sim-

ulated data, as well as human and animal participants, and used different actigraphy output by studying the counts per time interval rather than ENMO accelerometry data. Mathematically though, both types of observed relationships between subsampling interval and IV are possible, and the shape of this relationship will depend on the autocorrelation characteristics of the data, as well as the recording device used and the type of accelerometry output studied. In our case, the calculation of IV from a small sampling period would yield high positive correlation between adjacent sampling values of the time series, and this results in very low values of IV. The initial rapid rise in IV as the subsampling interval was increased from 5 seconds to higher multiple values can then be observed. In the right plot of Figure 5.2.3, we display the same results as the left plot, but this time the IV values from the intervention group and the group of individuals without dementia are represented as percentages of the values from the non-intervention group across the subsampling interval. Here we see that lower subsampling intervals were more effective at separating the groups. This indicates a clear trade-off in selecting the optimal subsampling interval.

To further study the appropriate choice of subsampling interval for IV, we investigated the effect of subsampling on the relationship between IV and our three other statistical measures. In Appendix A, we show plots of the Pearson's correlation coefficients, which were used to determine the linear association between IV and the other measures (see Figures A.2.1, A.2.2, and A.2.3). Because IV is used to assess the circadian fragmentation, then the measurement of strength of the circadian rhythm, as measured by IS or PoV, should have an associated negative correlation. Similarly, there should be a negative correlation between IV and the DFA scaling exponent,

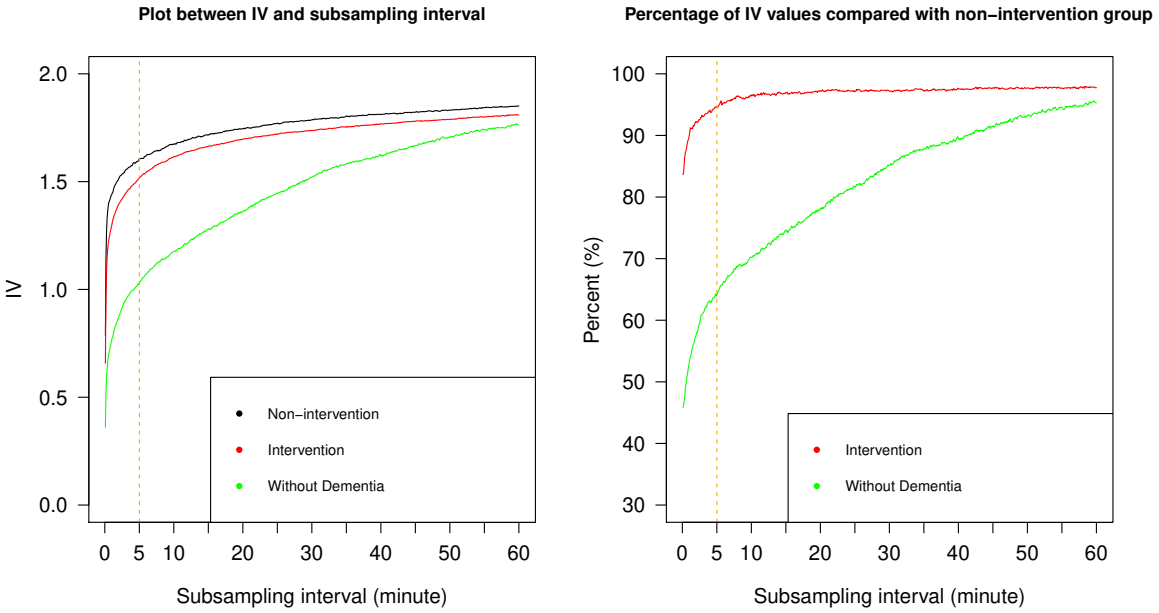

Figure 5.2.3: The left panel displays line plots of average IV values against the subsampling interval for the three different groups of individuals; the right panel displays line plots of the percentage of average IV values from the intervention group and the group of individuals without dementia, compared with those from the non-intervention group. The orange dashed line refers to 5 minutes and indicates our recommended subsampling interval.

as high IV and low DFA exponents are synonymous with rougher more volatile time series, whereas low IV and high DFA exponents correspond to smoother more meandering time series. The results in the supporting figures indeed revealed such negative correlations between IV and three other metrics, but only within suitable ranges of subsampling intervals that were not too high or too low, indicating that IV performed as expected in this range.

Taking together the evidence from Figures 5.2.3, A.2.1, A.2.2, and A.2.3, we sug-

gest selecting a subsampling interval in the range from 2 to 10 minutes. We selected 5 minutes when reporting our results that follow, as indicated by the dashed line in Figure 5.2.3 (although findings were broadly consistent selecting any value in this range). The range of 2 to 10 minutes partially overlaps with the findings of Gonçalves et al. (2014), where the most significant difference between groups of participants with and without dementia occurred when the subsampling rate was between 5 and 45 minutes. As mentioned, discrepancies between findings are likely due to the type of study being performed, as well as the specific recording device used, and we recommend such analyses are repeated in future accelerometry studies to ascertain the most appropriate subsampling rate which is an important choice for IV to perform meaningfully as a summary statistic.

Using a subsampling interval of 5 minutes, the corresponding IV value was calculated for each individual, as presented in Figure 5.2.4. Participants from both non-intervention and intervention groups usually had higher IV values than participants without dementia, as expected. As was the case with IS values, the participants without dementia had a larger spread of IV values than participants with advanced dementia. This was likely due to participants without dementia not having similar living conditions to each other, unlike participants with dementia who all resided in care homes. The mean IV value from the group of individuals without dementia was less than that from the combined non-intervention and intervention group (p-value < 0.001), but there was no significant difference in the mean IV value between these two groups of individuals with advanced dementia (p-value $= 0.357$).

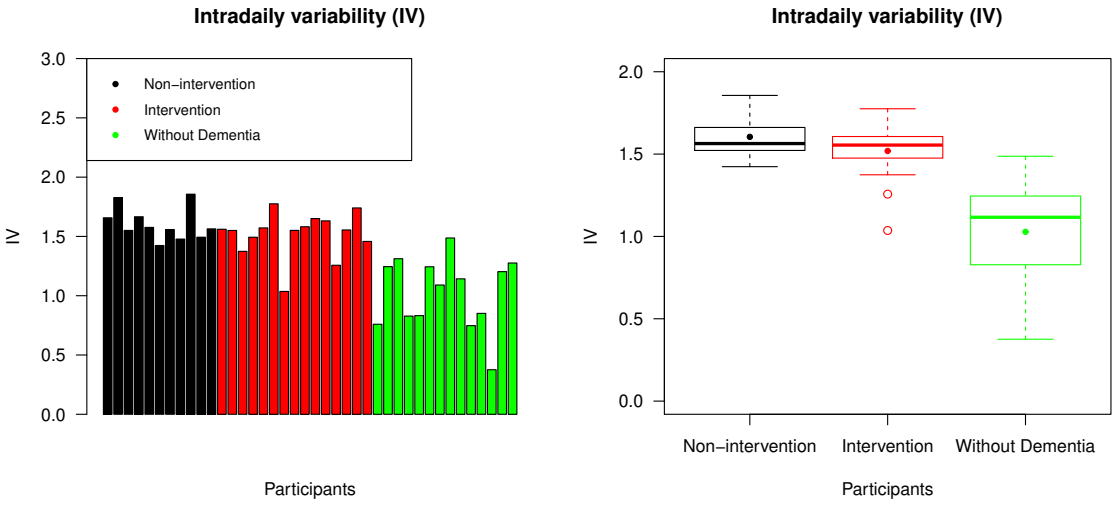

Figure 5.2.4: The left panel displays IV values across all participants, with the same ordering of individuals as the bar chart representing IS values; the right panel is a box plot collating IV values across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group, the median and mean of the IV values are presented by a thick line and a point in its own box, respectively.

5.2.4 DFA scaling exponent

The DFA scaling exponent (α) from each individual was estimated by the slope of the linear regression line between scale and overall root-mean-square deviation in log-log coordinates, as described in the Materials and Methods section. In Figure 5.2.5, the regression lines from three individuals from the different groups are presented. We evaluated the coefficient of determination (R^2) as a measurement of goodness of fit and found that approximately 99% of the variability of the data can be explained by the linear regression model in each case. From this figure, we can see that the participant without dementia has a higher slope, and hence higher α value, than the

participants from the non-intervention and intervention groups. We found this figure to be generally representative of the differences between DFA scaling exponents across all participant groups.

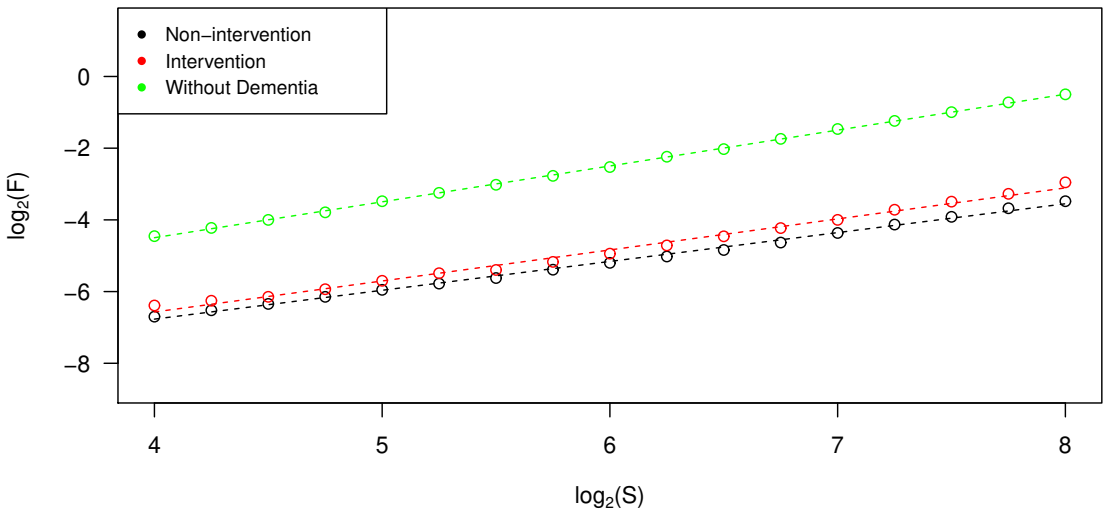

Figure 5.2.5: Data points and their linear regression lines for the estimation of DFA scaling exponents (slopes) from three participants, one from each of the non-intervention group, the intervention group, and the group of individuals without dementia.

Figure 5.2.6 shows DFA scaling exponents from all participants. The α values from the group of individuals without dementia were likely to be higher than those from the other groups (p-value < 0.001). In addition, participants from the intervention group had significantly higher mean α value than the non-intervention group (p-value $= 0.004$). All α values from these two groups with advanced dementia were between 0.7 and 1. This suggests that their corresponding ENMO time series were stationary

and noise-like but with long-term positive autocorrelations. For participants without dementia, α values were found to be very close to 1. The time series with these α values were in the boundary between stationary noise and nonstationary random walks and it was difficult to classify into either group. This type of time series data is sometimes called pink noise or $1/f$ noise. In general, the higher values of α observed in the group of individuals without dementia are consistent with time series that exhibited smoother and less jittery trajectories.

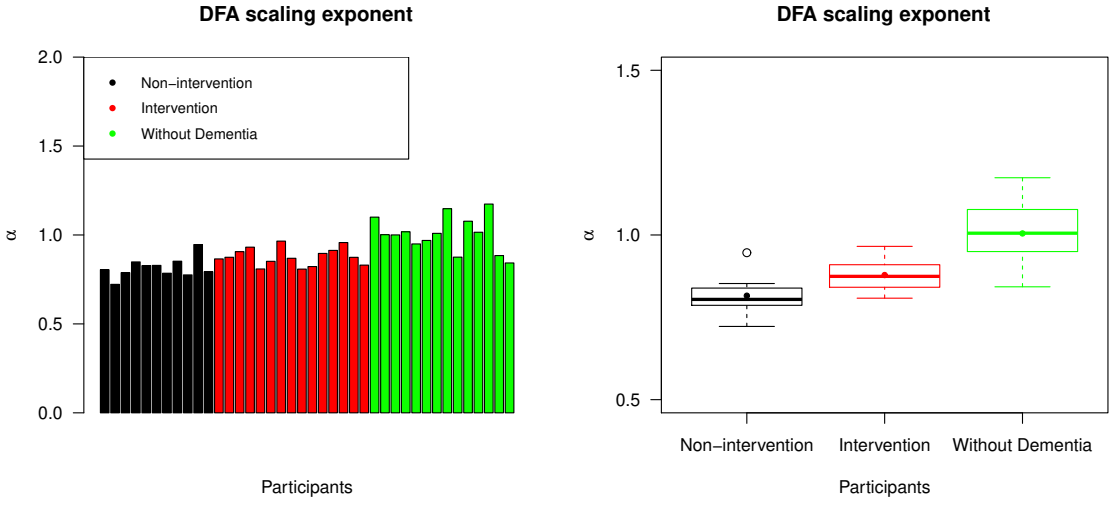

Figure 5.2.6: The left panel displays α values across all participants, with the same ordering of individuals as the bar chart representing IS values; the right panel is a box plot collating α values across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group, the median and mean of the α values are presented by a thick line and a point in its own box, respectively.

We also calculated separate DFA scaling exponents from daytime and nighttime readings of the ENMO data. Because each participant had a different sleeping period,

and there was no available data from the care homes regarding daily schedules, then setting fixed daytime and nighttime periods *a priori* was challenging. Instead we took a data-driven approach. Figure 5.2.7 displays the average EMMO readings throughout the day, averaged across each participant group and across all participants. In general, we observed low ENMO values between 11 p.m. and 6 a.m. and as such we used this seven hour interval as *nighttime* readings, and ENMO values between 6 a.m. and 11 p.m. as *daytime* readings.

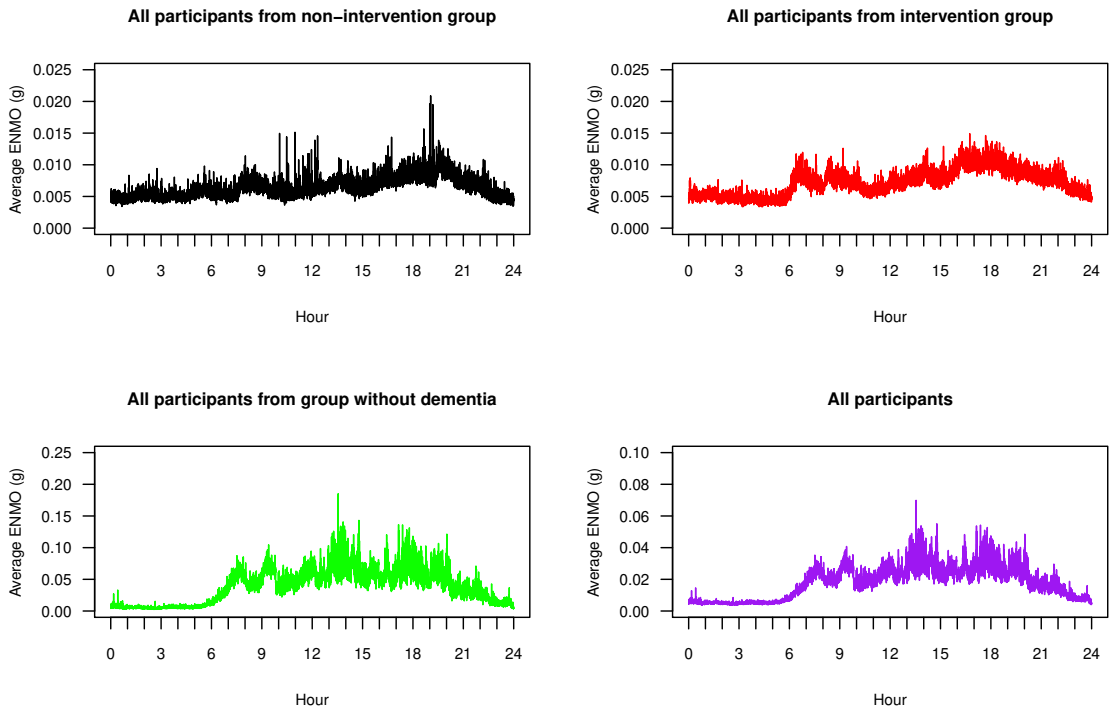

Figure 5.2.7: Time series plots of 24-hour ENMO values averaging over all participants in each of the non-intervention group (black), the intervention group (red), the group of individuals without dementia (green), and all groups together (purple).

Figure 5.2.8 shows two box plots presenting the summary statistics of separate DFA scaling exponents from daytime and nighttime readings. The daytime readings

had a similar distribution of α values to the entire time series analysed in Figure 5.2.6. The nighttime readings however provided a rather different distribution. Their α values were more similar across the groups and indeed there was no longer a significant difference of mean α value between participants in the intervention group and the group of individuals without dementia (p-value = 0.747). For the non-intervention and intervention group, there was no significant difference of mean α value between daytime and nighttime readings (p-values = 0.949 and 0.305). Participants without dementia, by contrast, had a significant difference of mean α value, with daytime values being higher (p-value = 0.008). These results suggested that participants with advanced dementia tended to have similar records of accelerometry throughout the day, and no significant change of fractal behaviour of data between daytime and nighttime. However, participants without dementia were likely to have this change of behaviour such that their accelerometry data could be separately described by two fractal exponents, one for daytime and one for nighttime.

5.2.5 Proportion of variance (PoV)

In Figure 5.2.9, we display the power spectral density, as estimated by the periodogram, from one participant from each of the non-intervention group, the intervention group, and the group of individuals without dementia, along with their corresponding percentage of the total variance at each frequency. The periodogram for the participant without dementia had its highest peak at a frequency of 1/24 hour^{-1}, which we call the first harmonic or fundamental frequency, and the next highest peaks were at the following higher orders of harmonics. These features were found in all

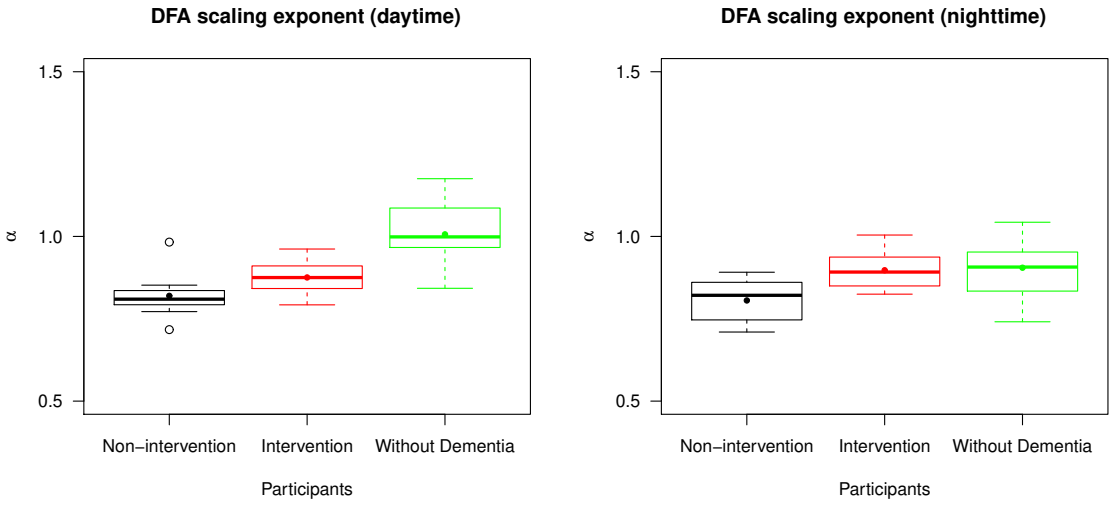

Figure 5.2.8: The left panel is a box plot collating α values from daytime readings and the right panel is a box plot collating α values from nighttime readings across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group and type of readings, the median and mean of the α values are presented by a thick line and a point in its own box, respectively.

other participants without dementia. Thus, their ENMO data was roughly periodic with a time period of approximately 24 hours or one day, corresponding to the time period of the circadian rhythm.

However, accelerometry data from individuals with advanced dementia was often distorted and this resulted in a much less clear pattern of circadian rhythm. In particular, the periodograms showed that the ENMO time series were less periodic, and the time period of the circadian rhythm was not well represented by the periodogram. In Figure 5.2.9, the periodogram for the participant in the non-intervention group did not show a clear and sharp peak at some of the harmonics including the funda-

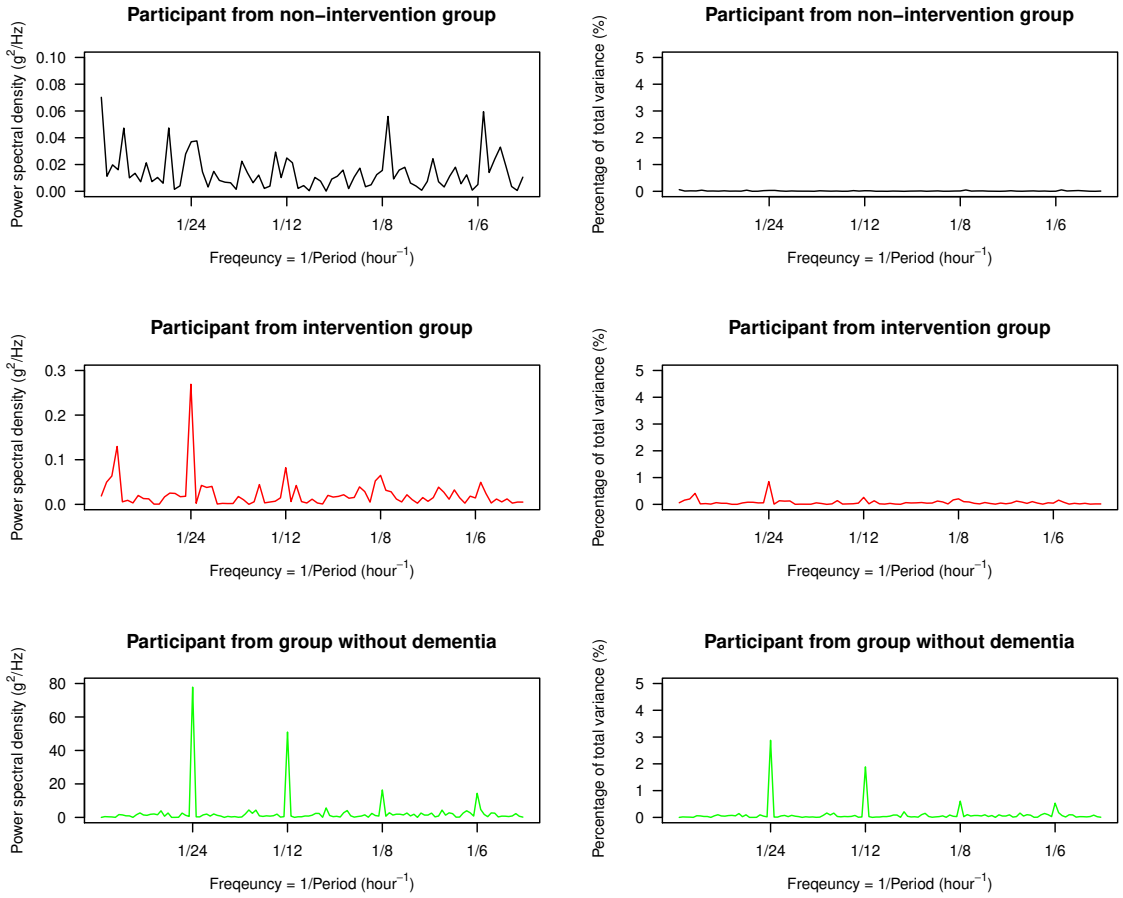

Figure 5.2.9: The left panel displays the plots of the power spectral density estimated by the periodogram for three participants, one from each of the non-intervention group, the intervention group, and the group of individuals without dementia. The right panel displays the plots of the percentage of the total variance captured at each frequency. The range of frequencies for each plot in both panels includes all four harmonic frequencies used in our study.

mental frequency. Although the highest peak could be observed at the fundamental frequency for the participant in the intervention group, there were less evident peaks at higher orders of harmonics and some peaks are located outside these frequencies.

These characteristics were generally representative of the participants from each respective group, however we note that there was variability across participants with advanced dementia in terms of the location of dominant peaks in the spectrum—the idea behind our PoV statistic is to smooth over such noisy characteristics to obtain a stable metric.

First we used Equation (5.1.9) to calculate $\text{PoV}^{(F)}$, which is the ratio of area under the spectral line around the fundamental frequency to the total variance of the time series (multiplied by 2 to include the negative frequency). This was used to indicate how well the variance was captured by periodic waves with period of $23.5 - 24.5$ hours. The plots in Figure 5.2.10 indicate that participants without dementia had significantly higher $\text{PoV}^{(F)}$ values than participants from the other two groups (p-value < 0.001). Overall, we see that this metric explained on average 15.161% of the variance for participants without dementia but typically less than 10% for participants with advanced dementia. We note that there was no significant difference in the $\text{PoV}^{(F)}$ values between intervention and non-intervention groups (p-value $= 0.217$).

To try and explain more of the variability, we then calculated $\text{PoV}^{(H)}$ from Equation (5.1.10) which included all periodogram values around the first four harmonic frequencies. The results are presented in Figure 5.2.11. Participants without dementia again had significantly higher PoV values than those with advanced dementia (p-value < 0.001), whereas there was no significant difference between intervention and non-intervention groups (p-value $= 0.384$). More variability was captured by this metric, in particular in the group of individuals without dementia where an average of 22.418% of the variance was captured.

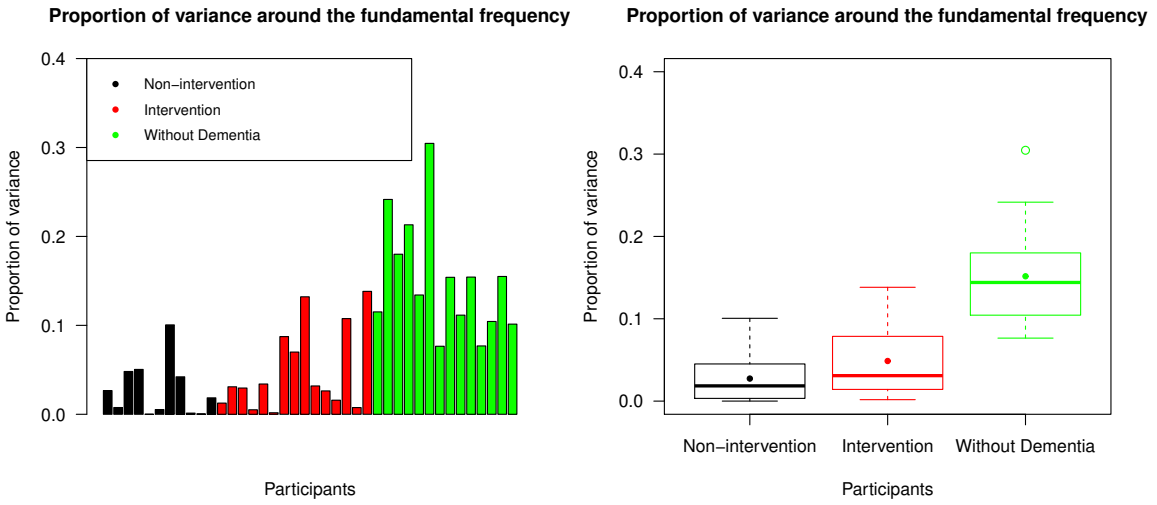

Figure 5.2.10: The left panel displays PoV values explained by the periodogram around the fundamental frequency across all participants, with the same ordering of individuals as the bar chart representing IS values; the right panel is a box plot collating these PoV values across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group, the median and mean of the $\text{PoV}^{(F)}$ values are presented by a thick line and a point in its own box, respectively.

Our proposed method of the proportion of variance (PoV) conceptually shares some similarities with the well-known cosinor method in that both methods measure the strength of periodic waves existing in the data. Indeed, we found that the coefficient of determination of the cosinor model (with 24-hour period) applied to our accelerometry time series was very highly positively correlated with $\text{PoV}^{(F)}$ across all participants ($\rho = 0.977$). This correlation was still strong but dropped to 0.901 when $\text{PoV}^{(H)}$ was used.

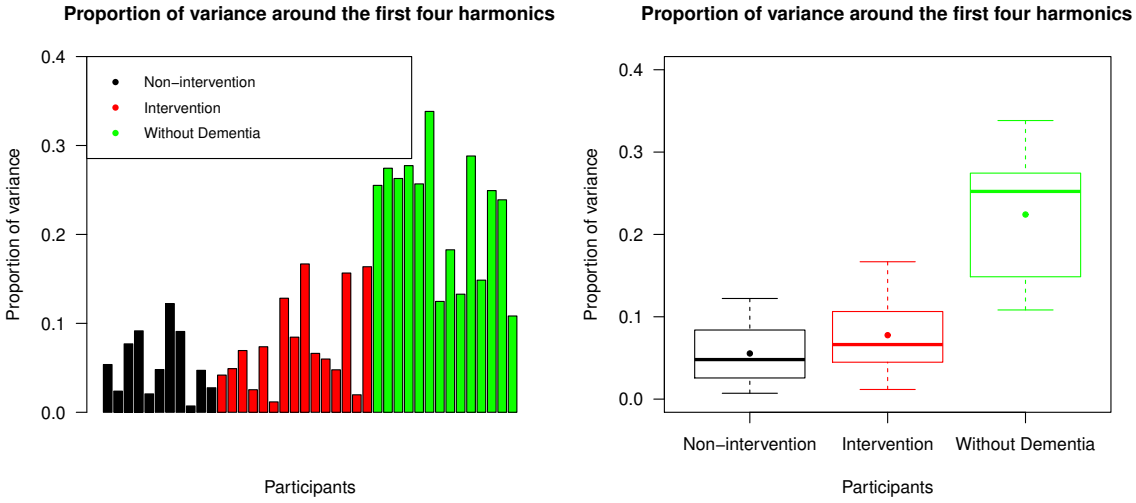

Figure 5.2.11: The left panel displays PoV values explained by the periodogram around the first four harmonic frequencies across all participants, with the same ordering of individuals as the bar chart representing IS values; the right panel is a box plot collating these PoV values across the non-intervention group, the intervention group, and the group of individuals without dementia. For each group, the median and mean of the $\text{PoV}^{(H)}$ values are presented by a thick line and a point in its own box, respectively.

5.2.6 Comparison of statistical measures

This section explores the relationships of results from all four proposed nonparametric methods. Scatter plots between all possible pairs of the four statistical measures—IS, IV (using a subsampling interval of 5 minutes), the DFA scaling exponent or α (from the whole range of the data), and $\text{PoV}^{(H)}$—are shown in Figure 5.2.12. We see that the statistics of participants from intervention and non-intervention groups were largely non-separable by any pair of these measures. However, the statistics of

participants without dementia were mostly distinct from participants with advanced dementia. Participants without dementia tended to have higher IS, α, PoV$^{(H)}$, but lower IV.

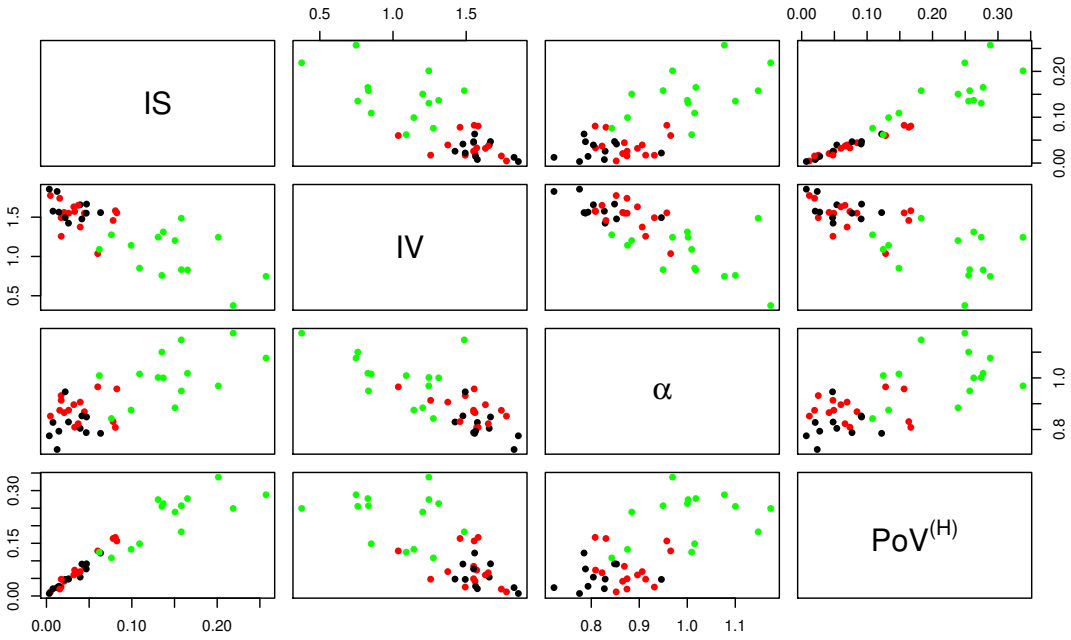

Figure 5.2.12: Scatter plots showing relationships between all possible pairs of the four statistical measures—IS, IV, the DFA scaling exponent (α), and PoV$^{(H)}$. Participants from the non-intervention group, the intervention group, and the group of individuals without dementia, are represented by black, red and green dots respectively.

In Table 5.2.1, we show the Pearson's correlation coefficients across all participants for all four measures. IV was negatively associated with all other measures, and all other pairwise associations were positively correlated. Such associations across groups were expected given the clear differences between accelerometry time series of participants without dementia and those with advanced dementia. Therefore, to explore

relationships between groups, we also separately summarised the Pearson's correlation coefficients for participants with and without advanced dementia in Tables 5.2.2 and 5.2.3, respectively. From these tables we can see that the relationships between all pairs of measures remained the same within these groups, albeit with lower correlations in general. This demonstrates that these statistical measures have power in separating within group behaviour and are useful statistics that should be considered when performing larger studies, for example to see if certain treatment interventions are effective when circadian rhythms are used as primary outcome.

Table 5.2.1: Pearson's correlation coefficients between pairs of statistical measures for all participants.

	IS	IV	α	PoV$^{(H)}$
IS		-0.775	0.735	0.943
IV	-0.775		-0.766	-0.724
α	0.735	-0.766		0.670
PoV$^{(H)}$	0.943	-0.724	0.670	

Table 5.2.2: Pearson's correlation coefficients between pairs of statistical measures for participants with advanced dementia (combined non-intervention and intervention groups).

	IS	IV	α	PoV$^{(H)}$
IS		-0.348	0.148	0.983
IV	-0.348		-0.600	-0.406
α	0.148	-0.600		0.169
PoV$^{(H)}$	0.983	-0.406	0.169	

Table 5.2.3: Pearson's correlation coefficients between pairs of statistical measures for participants without dementia.

	IS	IV	α	PoV$^{(H)}$
IS		-0.434	0.509	0.772
IV	-0.434		-0.441	-0.220
α	0.509	-0.441		0.308
PoV$^{(H)}$	0.772	-0.220	0.308	

5.3 Conclusions

This chapter proposes four nonparametric summary statistics (IS, IV, the DFA scaling exponent, and PoV) for the study of circadian rhythm and other attributes found in accelerometry data. As a proof-of-concept, we computed each summary statistic on a dataset containing a mixed group of participants, some with advanced dementia, and compared these results with data from individuals without dementia. The analysis shows that these statistics can collectively summarise different features in the data: both the inter-daily features of the circadian cycle (IS and PoV), as well as the intra-daily fragmentation and correlation structure (IV and the DFA scaling exponent). Using these statistics there are clearly different values obtained for participants without dementia and those with advanced dementia, therefore there is a potential for these nonparametric statistics to be used as primary outcomes related with circadian rhythms, or as diagnostic benchmarks and thresholds, which would naturally require more studies and analyses to precisely establish practical clinical guidelines.

Some of the participants with advanced dementia also received a group intervention during the study. As in the original study (Froggatt et al., 2020), we did not observe statistically significant differences between the intervention and non-intervention groups. There was however one exception presented here, where the DFA scaling exponent values (whether daytime, nighttime, or aggregated across both) did show significant difference between non-intervention and intervention groups. Specifically, the scaling exponent was found to be significantly higher for the intervention group which indicated a smoother pattern of behaviour in activity over longer time scales, which was a feature we also found in the group of participants without dementia. Overall, the fact that participants with and without dementia displayed clear differences to each other, but participants within the groups with advanced dementia did not, were both expected results and validated our statistical measures in terms of their efficacy and reliability in summarising key features of accelerometry data. Ultimately to determine the scope for using these summary statistics in a clinical trial setting requires much further study.

The accelerometry data we studied is high frequency—sampled every 5 seconds—and this was typical of modern instruments and studies. We paid careful consideration as to how statistical methods should be adapted to high frequency sampling. A key finding is that IV should *not* be calculated from hourly subsampling, as has been historically performed in the literature. Hourly subsampling leads to IV values that are unable to effectively separate groups of participants, and have unexpected correlations with other summary statistics. Instead, we proposed subsampling every 5 minutes for our studied dataset, which improved the behaviour and increased the

power of this statistic. We note that subsampling at even higher frequencies eventually worsened performance due to contamination from high-frequency variability. Finally, we note that we proposed a simple aggregation method to ensure no data is lost in the subsampling procedure.

The high frequency nature of the data also allowed us to propose a new metric, *proportion of variance* (PoV), which is a nonparametric spectral-based estimate of circadian strength. As the data is high frequency, there are many frequencies at or near the circadian frequency which can be smoothed over to obtain a stable estimate of circadian strength. A key finding is that the inclusion of *harmonics*, which are integer multiples of the circadian frequency, increased the proportion of variance explained by this statistic and led to improved performance. This statistic is therefore a nonparametric alternative to the parametric cosinor method. We note however that PoV performs similarly to IS with high correlation in these values across participants, and it is therefore possible that only one of these metrics is required, which future studies may reveal.

Finally, the high frequency nature of our data allows for statistics to be calculated to quantify the *memory* or *long-term autocorrelation* structure of the time series. A variety of established techniques exist to quantify such structure, including the DFA method used here. What our analysis revealed however, is the potential for utilising the richness of continuity of accelerometry data to obtain separate daytime and nighttime DFA scaling exponent values, and how this may yield further insight as to the difference between individuals who have disrupted sleep cycles and activity rhythms, and those that do not. In particular, we expect individuals without dementia

to have significantly higher DFA scaling exponents at daytime than at nighttime, and this is what our analysis revealed.

In this chapter, we consider three existing nonparametric metrics—IS, IV and the DFA scaling exponent—for analysing accelerometry data, together with proposing a novel nonparametric alternative to cosinor analysis based on aggregating information in the periodogram. We provide guidelines on implementation, and an indication of their performance on a high frequency dataset from people suffering from advanced dementia living in care homes. Limitations of our data analysis include having a relatively small number of participants, and that the without dementia group data comes from individuals from the research group with younger ages than those with advanced dementia living in care homes. Different demographic variables such as age will affect the interpretation of statistical results from accelerometry data and hence, at this point, we cannot conclude that the difference of statistical measures between participants in our case study is solely due to the condition of having advanced dementia or not. Therefore care should be taken in terms of generalising findings from our data analysis, and the performance and robustness of proposed summary statistics should be rechecked in future accelerometry analyses. More broadly, high frequency accelerometry analysis is a relatively recent development in the health sciences, and we very much encourage continued exploration and refinement of statistical methodology as the complexity and dimensionality of acquired datasets continue to grow.

There are some possible future works that could be discussed here. In our work, we excluded all data for any day having at least one missing data, so the data after this removal have no missing values, and maintain the pattern of daily cycle with 24-

hour period. However, such removal may result in a loss or degradation of statistical properties given by several functions such as the autocorrelation. The analysis from original data with some missing values may provide different results of our nonparametric measures which should be considered. Another future work is to perform the calculation of the DFA scaling exponent varying with the local temporal fluctuation of accelerometry time series. This can be performed by the multifractal detrended fluctuation analysis (MFDFA) (see Ihlen (2012)). Different values of the DFA scaling exponent varying in time can be derived from each time series using this method, to further separate differences between groups of individuals.

Chapter 6

The Study of Several Measures to Detect Fatigue in Sport Sciences

In sports science, *fatigue* is defined as the decrease or failure in the muscles' ability to have their full performance in a game or an exercise due to the change in biochemical substances in the players' muscles (Edwards, 1983). The study of fatigue from professional athletes has been greatly increased in the past few decades, in part thanks to the development of technological and innovation systems to frequently and accurately detect players' movements with respect to time, and measure the change of biochemical compounds in the muscles over time. It is one of the crucial research topics in professional association football (soccer) as fatigue inevitably has impact on both individual and team performances in a match, and perhaps is a key factor of the final result of the game. Statistical analysis has been applied to either physical (Mohr et al., 2003; Di Salvo et al., 2009; Rampinini et al., 2009) or biochemical (Krustrp et al., 2006; Bangsbo et al., 2007; Rampinini et al., 2011; Silva et al., 2018) measures

related to fatigue. The main goal in this chapter is to perform detailed statistical analysis including calculating physical measures such as the rate of change of distance covered by a player in a match, and the variability of players' acceleration data. In particular, we will check the ability of these measures to detect the impact of fatigue on the performance of footballers in different playing positions.

Fatigue usually occurs during a highly intensified period, and has a huge effect on footballers' performance towards the end of the game. Several changes of their performance have been reported such as the significant decrease of (1) the amount of high-intensity running and sprinting and (2) the total distance covered at these ranges of speed from non-substituted players (Reilly, 1997; Mohr et al., 2003, 2005). The comparison of individual performance from footballers in different playing positions has also been studied. Reilly (1997) stated that the total distance covered was related to the energy expenditure which was the highest among midfielders according to their findings. Mohr et al. (2003) found that the mean of the total distance covered from defenders was significantly lower than all other outfield players. They also conducted the research on the measurement of temporary fatigue by dividing a whole period of the game into a collection of non-overlapping 5-minute periods, and found a substantial decrease of the distance covered at high-intensity running during the 5-minute period after the peak period with the highest distance covered such that only 88% of the mean 5-minute distance covered was covered in this period. Akenhead et al. (2013) further investigated the temporary fatigue by measuring the total distance covered at high acceleration and high deceleration, and they found similar results in this 5-minute period after the peak period to those from the high-intensity running

in Mohr et al. (2003).

The measurement of total distance covered is usually derived from equipment that can detect the movement patterns of players. The individual video recording of the players with VHS-format cameras was used in Mohr et al. (2003). However, this is no longer used in the present day due to its time-consuming nature and technological advances now available. Several methods have been developed to replace this old-fashioned approach for tracking players in the field such as semi-automatic passive multiple-camera systems (MCS) (Bradley et al., 2009; Di Salvo et al., 2009) and automatic GPS (Akenhead et al., 2013; Oliva-Lozano et al., 2020). A more advanced system with a specialised GPS receiver, a laser, and a scanner was developed to generate the three dimensional representations of the location on the surface by the reflection of ultraviolet or visible light, and this system is known as LIDAR (Light Detection and Ranging). This is a very new approach for tracking footballers developed by Sportlight Technology Ltd. (https://www.sportlight.ai), and at the time of writing there have been no published scientific papers about this system for sport applications such as football. The high-frequency tracking data from a LIDAR system can be used to acquire information of distance, speed, and acceleration of each player in the field at high frequency (around 10 Hz).

Levels of fatigue have been studied in many sports and events. These include several major leagues of association football, for example, the UEFA Champions League (Di Salvo et al., 2010), the English Premier league (Bradley et al., 2009; Di Salvo et al., 2009; Akenhead et al., 2013), the Spanish LaLiga (Oliva-Lozano et al., 2020), and the Italian Serie A league (Rampinini et al., 2009). Some other sports

with research evidence on the study of fatigue are rugby (Waldron et al., 2013; Twist and Highton, 2013), basketball (Lyons et al., 2006; Edwards et al., 2018), and baseball (Mullaney et al., 2005). These studies are important to learn how performance of athletes changes over time and what optimal strategies should be applied to make the best outcome for both individual and team performances due to fatigue, as well as their impact on injury.

The details on measures related to fatigue in this chapter are provided as follows. The first section, *Materials and Data*, summarises all features of the high-frequency tracking data from the LIDAR system. All data were collected from footballers of one English Premier League team during the 2020/2021 season. The details of the team's name, match dates, and players' names are omitted due to data sensitivity and privacy considerations. Several criteria are set to select suitable data for our statistical analysis. The second section, *Methodology*, explains the methods of statistical measures used for studying the impact of fatigue on the performance of players in different playing positions. First, we measure the total distance covered at high *acceleration* and high *deceleration*, and this is similar to the study in Akenhead et al. (2013). Next, a new feature of data called the *significant turn* is introduced. Specifically, a significant turn is recorded when a player changes their movements from very high deceleration to very high acceleration in a short amount of time. A new statistical measure comparing performance of players before and after each significant turn will be given. Furthermore, the process for constructing a regression model of count data of significant turns is explained. Lastly, two nonparametric measures including the intradaily variability (IV) and the scaling exponent (α) from detrended fluctua-

tion analysis (DFA) are estimated from the acceleration time series of high-frequency tracking data. Although these measures are common in biomedical sciences, they will be used to check their effectiveness and potential for acceleration data from footballers in this chapter. We are not aware of other studies which have attempted such an analysis. The third section, *Results and Discussion*, provides the calculation of all statistical measures given in the Methodology section. The summary statistics and their comparison between footballers in different playing positions or different periods of the game will be discussed.

6.1 Materials and Data

All high-frequency tracking data were recorded from LIDAR systems developed by Sportlight Technology Ltd. They collaborated with an English Premier league team in a single season to record players' movements in all matches played at their home ground. The data were provided with two sets as follows:

The first dataset

1. Timestamps at every 0.1 seconds of the game (sampling period of data).
2. xy-coordinates of positions of players away from the location of the sensor.
3. Angles or directions of movements (in degrees).
4. Smoothed speed (in metres per second or m/s) and smoothed acceleration (in metres per second squared or m/s^2) calculated through a specific low pass filter bidirectionally.

The second dataset

1. Dates.
2. Players' ID numbers.
3. Timestamps when significant turns started to occur.
4. Direction of significant turns (either left or right).
5. Angles of significant turns (in degrees).
6. Peak values of deceleration and acceleration during significant turns.
7. Starting and ending speeds during significant turns.
8. Times between the ending of high deceleration and the starting of high acceleration during significant turns (twisting times).

Because all personal data were confidential and unable to be accessed, players' names were replaced by the unique ID numbers, which were randomly changed in every match. Hence, the demographic characteristics of all players are unknown, and this contributes to some limitations of our data analysis. In addition, no categorical data of their playing positions were given. However, we estimated these playing positions by inspecting their average movements in each half of the match. Figure 6.1.1 shows examples of movements in the first half (black line) and the second half (green line) from three players. The heat maps for the distribution of their movements only in the first half are also reported in Figure 6.1.2. According to these two figures, we observe different high-density locations from these players, and this information could be used

to predict their playing positions, as labelled above each figure. It is of note that the location of the sensor is at the origin $(0, 0)$. Because it was set at a high level above and not exactly in the middle between both ends of the field, the two-dimensional mapping of the field does not align with both x and y axes. We also calculated the mean of all coordinates of positions of player in each half. The visualisation of all means from players in a single match in Figure 6.1.3 could be used to classify their playing positions.

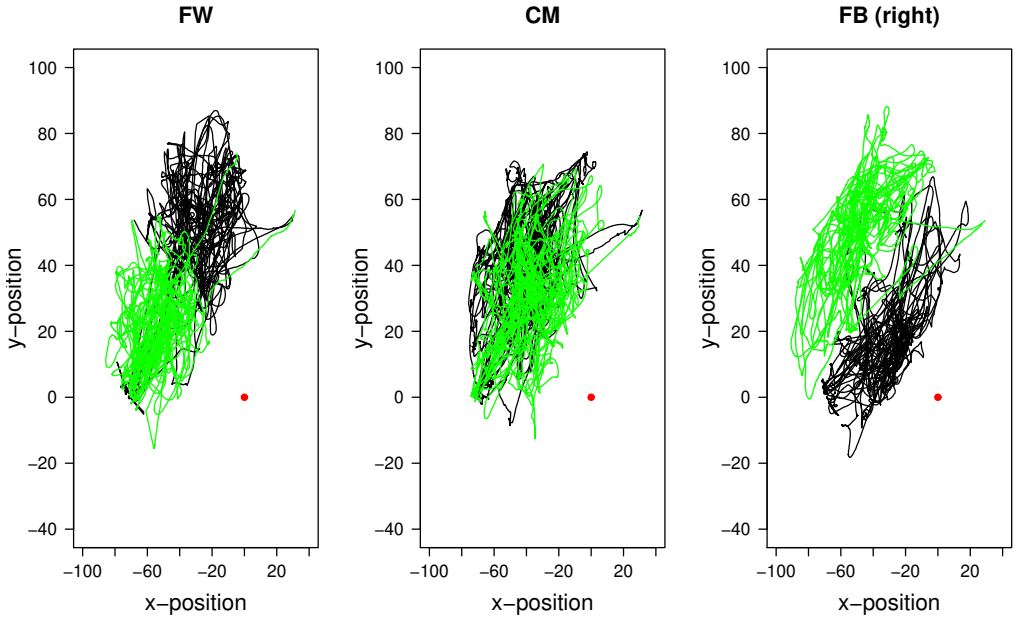

Figure 6.1.1: Line plots of movements from three different players in the first half (black line) and the second half (green line) of a single match with the sampling period of 0.1 seconds. A red dot in each plot shows the location of the sensor used in the LIDAR system which was developed by Sportlight Technology Ltd.

As we observed either a $4-4-2$ or $4-4-1-1$ formation in all matches played by the team in our study, six different groups of playing positions were assigned including the

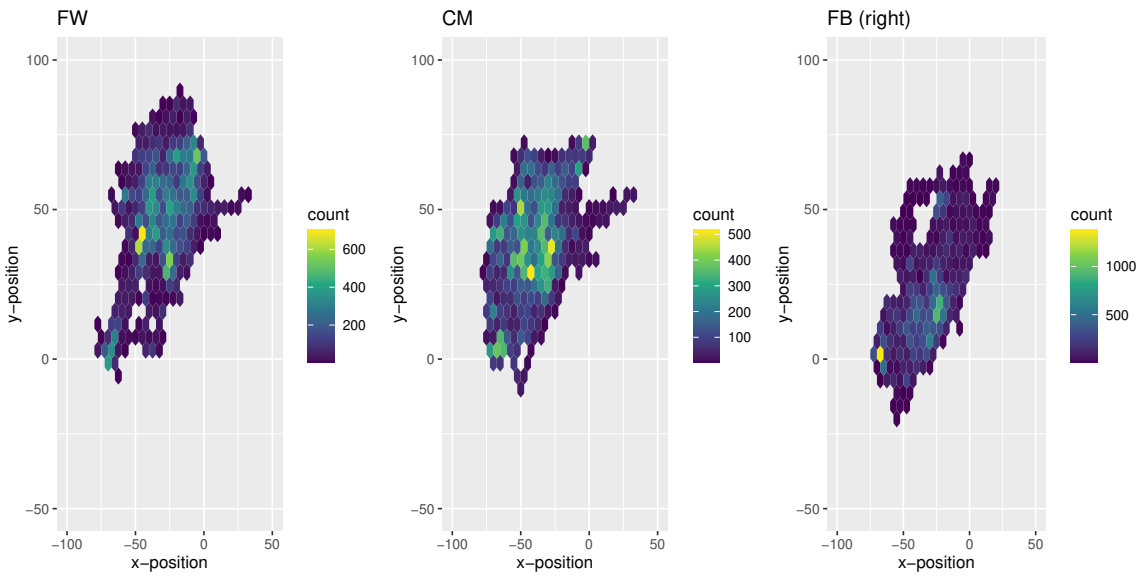

Figure 6.1.2: Heat maps for the distribution of players' movements in the first half shown in Figure 6.1.1 (black line).

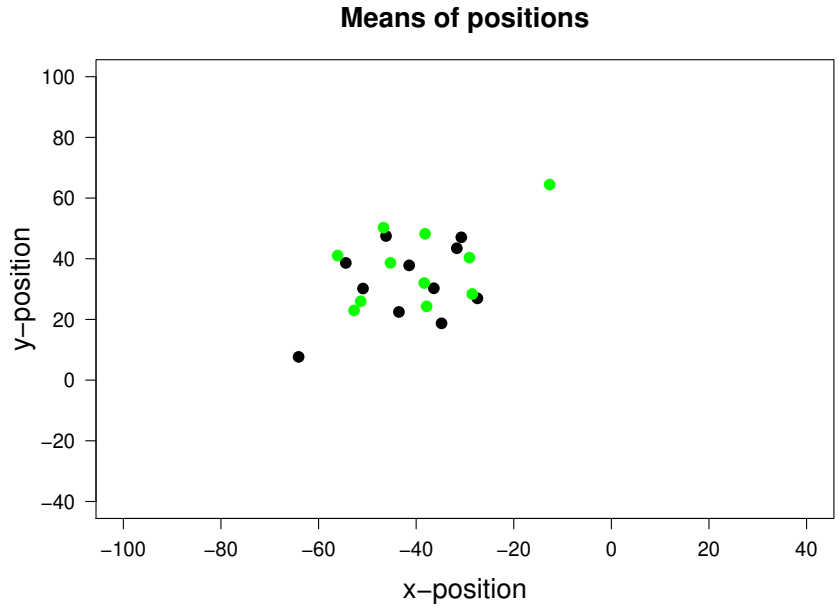

Figure 6.1.3: Points of means of positions of all eleven players in the first half (black) and the second half (green) of a single match with no substitution.

forwards (FW), the wide midfielders or wingers (WM), the central midfielders (CM), the full-backs (FB), the centre-backs (CB), and the goalkeeper (GK). Each group except the goalkeeper consists of two players per time in a single match. Some of the available datasets from players in different playing positions were excluded from our statistical analysis, which was performed in RStudio with R version 4.0.0, if they did not meet one of these following criteria:

1. There was no consecutive missing data for more than one minute of a game.
2. Data were from players who played in the full match with no substitution.
3. Data were from players who played in the same position throughout the match.
4. Data were from players who did not have any serious injury during the match.

For these reasons, we excluded 33 out of 55 datasets of FW, 25 out of 49 datasets of WM, 10 out of 42 datasets of CM, 5 out of 40 datasets of FB, and 3 out of 39 datasets of CB. These omitted datasets are useful for further study to assess the impact of substitutions and injury and how this relates to fatigue, but this is reserved for our future work. Instead we focus on complete datasets for fair comparison of fatigue across the game and across playing positions.

6.2 Methodology

In this section, we detail the methodology used to analyse the data. In Section 6.3, we will report and discuss the results found.

6.2.1 Distance Covered

We calculated the total distance covered from xy-coordinates of positions of a player in each half of a match. This was approximated by the summation of all Euclidean distances derived from each of two consecutive points measured every 0.1 seconds. Akenhead et al. (2013) introduced the study of impact of fatigue on players' performance with the calculation of the total distance covered at different ranges of speed and acceleration. In this chapter, we have repeated the analysis from Akenhead et al. (2013) using our dataset. The ranges of speeds and accelerations we considered include

- High speed (HS: ≥ 5.8 m/s)
- Sprint (SP: ≥ 6.7 m/s)
- Total acceleration (T_{ACC}: ≥ 1 m/s^2)
- Low acceleration (L_{ACC}: between 1 and 2 m/s^2)
- Moderate acceleration (M_{ACC}: between 2 and 3 m/s^2)
- High acceleration (H_{ACC}: > 3 m/s^2)
- Total deceleration (T_{DEC}: ≤ -1 m/s^2)
- Low deceleration (L_{DEC}: between -2 and -1 m/s^2)
- Moderate deceleration (M_{DEC}: between -3 and -2 m/s^2)
- High deceleration (H_{DEC}: < -3 m/s^2)

We calculated the summary statistics (mean and standard deviation) of the total distance covered from all players in both halves and the full match excluding injury time at each range of speed and acceleration. In addition, these results at high acceleration in magnitude (a combined range of H_{ACC} and H_{DEC}) from outfield players in different playing positions were analysed. The comparison of the total distance covered from players at each range of speed and acceleration between both halves was conducted using a paired t-test with a significance level of 0.05. This test could be performed as the total distance covered is likely to be normally distributed with p-value greater than 0.05, according to the Shapiro-Wilk test. Such significance results were also reported in Akenhead et al. (2013).

The total distance covered at the combined range of H_{ACC} and H_{DEC} was further investigated in non-overlapping six 15-minute periods of the game. The means of the total distance covered between each pair of groups of players were compared using the one-way ANOVA with the Tukey-Kramer post-hoc test, which is a robust method under the scenario of unequal sizes of samples and large number of groups of samples with no specific planned comparison. This test with a significance level of 0.05 was implemented in R using the TukeyHSD() function.

Lastly, we calculated the total distance covered at high acceleration in magnitude in four defined periods according to Akenhead et al. (2013). One of these periods was the 5-minute peak period or 5_{PEAK}, which is the period when a player had their highest total distance covered at H_{ACC} or H_{DEC} in a single match. The others include 5-minute period before (5_{PRE}), 5-minute period after (5_{POST}), and 10-minute period after (10_{POST}) the peak period. The relative percentage of the total distance covered

in each period to the overall mean in the half when the peak period occurred was calculated and compared. Unlike Akenhead et al. (2013), this overall mean excludes the total distance covered during the peak period or 5_{PEAK}, which we argue is a fairer comparison to detect fatigue using a relative measure, as we shall discuss further.

6.2.2 Significant Turns

We seek to analyse the impact that significant turns have on fatigue. Therefore, for this chapter, we introduced a new statistical measure which is the ratio of the distance covered at high acceleration in magnitude to the total distance, calculated both before and after a significant turn. Professional footballers normally try to reduce their risk of fatigue by avoiding having high acceleration or high deceleration of their movements unless necessary. Alternatively, a player may simply be fatigued by a significant turn and other events in the game and their high-intensity output levels drop in the proceeding period of time. Either way, after each significant turn, we might expect our measure (the ratio of the distance covered at high acceleration in magnitude to the total distance) to decrease as players are fatigued or seek to avoid further fatigue. The duration of each period before or after the significant turn was set to one minute. If the time difference between two significant turns is less than one minute, the time duration of the period after the first turn and the period before the second turn was set to that time difference instead. In Appendix B, we provide a discussion on changing the duration of the period before and after the significant turn from between one second to two minutes. The comparison of this measure between the periods before and after was conducted using the one-sided paired t-test with a

significance level of 0.05. Furthermore, the summary statistics of this measure were calculated from all and each group of outfield players.

The total number of significant turns from each player was counted and compared to others with different playing positions. This was also conducted in both halves and the full match as we also performed with the total distance covered. We found that the expected number of significant turns could be predicted by a regression model using the total distance covered and the position as its explanatory variables (covariates). The total number of significant turns could be treated as count data, and the most common model for this type of data is the Poisson regression model. Therefore we used a generalised linear regression model having the logarithm as its canonical function. The Poisson distribution for this model is given by

$$\begin{aligned} f_Y(y) &= \frac{\lambda^y \exp(-\lambda)}{y!} \\ &= \exp(y \log \lambda - \lambda - \log(y!)), \end{aligned} \tag{6.2.1}$$

which is a member of the exponential family. The realisation of a non-negative count Y is a response variable y of the total number of significant turns, and $\mathbb{E}(Y|\mathbf{X}) = \mathrm{Var}(Y|\mathbf{X}) = \lambda$, where $\mathbf{X}$ is a vector of explanatory variables of the model. The canonical link $g(\lambda) = \log \lambda$ is used as the linear predictor of the model such that $\log \lambda = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$, where $\{X_i\}_{i=1}^{n} \in \mathbf{X}$ is a set of n explanatory variables and $\{\beta_i\}_{i=0}^{n} \in \mathbf{B}$ is a set of $n+1$ model parameters. Because the position is a factor variable, it is worth checking whether there is a different relationship between the total number of significant turns and the total distance covered at each level of position. This relationship is known as an interaction resulting in the multiplicative effects of

explanatory variables in the model. The chi-square difference test was used for the model selection among three purposed quasi-Poisson regression models, which take an account of the over-dispersion problem due to a huge difference between the mean and the variance of count data, as follows:

- Model I: $\log \lambda = \beta_0 + \beta_1 D + \sum_{i=1}^{4} \beta_{i+1} G_i + \sum_{i=1}^{4} \beta_{i+5} D G_i$,
- Model II: $\log \lambda = \beta_0 + \beta_1 D + \sum_{i=1}^{4} \beta_{i+1} G_i$,
- Model III: $\log \lambda = \beta_0 + \beta_1 D$,

where D is the continuous variable of the total distance covered, and G_i is the indicator function for the i^{th} group of outfield player, where the first group is "CM", the second group is "FB", the third group is "FW", and the fourth group is "WM". The base level such that $G_i = 0$ for all $i = 1, 2, 3, 4$ represents the group of "CB".

6.2.3 Nonparametric Measures

In the previous chapter, we reported several nonparametric measures for accelerometry data where a circadian rhythm was presented. Although the circadian pattern does not exist in the acceleration data of footballers within a 90-minute game, some of these measures (the ones not related to measuring the circadian rhythm) were selected to study the variability and long-memory behaviour of the footballers' acceleration data including intradaily variability (IV) and the scaling exponent (α) from DFA.

Intradaily variability (IV) was used to measure the degree of fluctuation or variability of data. The subsampling period for the calculation of IV was set to one second

so that it could capture the variation of acceleration in every second. Let $\{X_t\}_{t=1}^N$ be the subsampled data. The IV was derived from the ratio of variation in consecutive times over the subsampling interval to the total variation of the data, i.e.,

$$\text{IV} = \frac{\sum_{t=2}^{N}(X_t - X_{t-1})^2/(N-1)}{\sum_{t=1}^{N}(X_t - \bar{X})^2/N}. \tag{6.2.2}$$

Furthermore, if $\{X_t\}$ is stationary Gaussian white noise, its expected value of IV is given by

$$\mathbb{E}(\text{IV}) = \frac{\mathbb{E}(X_t^2) - 2\mathbb{E}(X_t X_{t-1}) + \mathbb{E}(X_{t-1}^2)}{\mathbb{E}(X_t^2) - \mathbb{E}(X_t)^2} = 2. \tag{6.2.3}$$

This reference value of 2 will help us to interpret the fractional behaviour of the time series and its potential relationship with FGN. Therefore, if $H = 0.5$ then IV $= 2$ as established in Equation (6.2.3). However, if $H > 0.5$ then there is a positive lag-1 autocorrelation and IV < 2. Conversely if $H < 0.5$ then there is a negative lag-1 autocorrelation and IV > 2.

Detrended fluctuation analysis (DFA) is another nonparametric method for the measurement of the degree of fluctuation or variability of data by using its scaling exponent (α). This was derived from the relationship of $F(S) \propto S^{\alpha}$, where $F(S)$ is the root-mean-square deviation from fitting a linear line to a segment of data with length S. More details on the estimation of α was already given in Section 5.1.4 of Chapter 5. A lower value of α indicates that the fluctuation is likely to be larger and more spikey. The time series is considered to be stationary if $\alpha < 1$, or nonstationary if $\alpha > 1$. The scaling exponent is related to the Hurst exponent by $H = \alpha$ for stationary data and $H = \alpha - 1$ for nonstationary data.

6.2.4 Comparison between Statistical Measures

The comparison between two nonparametric measures, IV and α, was conducted to ensure that their correlation results agreed with what we observed from the statistical analysis of accelerometry data in the previous chapter. In addition, we provided a comparison of these results between each nonparameteric measure and the total number of significant turns in the full match.

6.3 Results and Discussion

6.3.1 Distance Covered

Table 6.3.1 reports the mean and standard deviation of the total distance covered at each observed range of speed and acceleration from all outfield players in both halves and the full match. The means of the total distance covered in the second half are significantly lower than the first half at all ranges. This finding might be attributed to the energy reduction of players resulting in the increasing impact of tiredness or fatigue on players' performance, and is consistent with the findings of Reilly (1997); Mohr et al. (2005); Akenhead et al. (2013).

Next, Table 6.3.2 reports the mean and standard deviation of the total distance covered at high acceleration and high deceleration from each group of outfield players (by playing position) in both halves and the full match. The number of players in each group is different due to our criteria for the data selection. According to the paired t-test, their playing position is likely to have an effect on their total distance covered

Table 6.3.1: The mean and standard deviation (mean ± SD) of the total distance covered in metres at each range of speed and acceleration from all 149 datasets of outfield players in both halves and the full match.

Range	First half	Second half	Full match
All	5429.34 ± 501.22	5319.99 ± 427.61	10749.33 ± 848.25
HS	338.78 ± 123.46	308.12 ± 108.91	646.90 ± 204.30
SP	126.76 ± 68.73	111.80 ± 60.11	238.56 ± 107.42
T_{ACC}	806.85 ± 141.93	753.30 ± 124.02	1560.15 ± 230.57
L_{ACC}	559.02 ± 95.12	523.18 ± 82.53	1082.20 ± 154.01
M_{ACC}	174.71 ± 36.98	162.55 ± 32.97	337.26 ± 59.51
H_{ACC}	73.12 ± 20.59	67.56 ± 22.30	140.68 ± 34.67
T_{DEC}	707.29 ± 135.64	648.06 ± 115.87	1355.35 ± 219.86
L_{DEC}	485.20 ± 88.16	445.77 ± 73.83	930.97 ± 133.41
M_{DEC}	136.75 ± 33.60	123.93 ± 27.33	260.68 ± 54.70
H_{DEC}	85.34 ± 25.23	78.36 ± 25.78	163.70 ± 46.87
H_{ACC} & H_{DEC}	158.46 ± 42.59	145.92 ± 43.10	304.38 ± 75.83

throughout the game. In the full match, the wide and central midfielders have the highest means. This is likely due to their responsibility to control the balance between offensive and defensive plays requiring more distance covered than other players. The centre-backs, on the other hand, have the lowest mean as their main role is to primarily defend in the centre area in front of the goalkeeper. This area of players' movements is likely to be smaller in size than of other playing positions in the field.

Table 6.3.2: The mean and standard deviation (mean ± SD) of the total distance covered in metres from each group of outfield players at acceleration greater than 3 m/s^2 in magnitude in both halves and the full match.

Position (number of players)	First half	Second half	Full match
FW ($n = 22$)	153.23 ± 41.22	135.82 ± 40.40	289.05 ± 71.34
WM ($n = 24$)	189.42 ± 34.61	165.05 ± 51.57	354.47 ± 69.23
CM ($n = 32$)	195.82 ± 22.80	185.01 ± 23.85	380.83 ± 31.17
FB ($n = 35$)	145.48 ± 33.77	138.52 ± 29.95	284.00 ± 49.75
CB ($n = 36$)	120.44 ± 26.25	111.77 ± 28.07	232.21 ± 43.19

A regular period of 90 minutes in a match was divided into six non-overlapping 15-minute periods: P1 (1^{st} to 15^{th} minute), P2 (16^{th} to 30^{th} minute), P3 (31^{st} to 45^{th} minute), P4 (46^{th} to 60^{th} minute), P5 (61^{st} to 75^{th} minute), and P6 (76^{th} to 90^{th} minute). Note that data from injury time periods (after the 45^{th} minute in the first

half and after the 90^{th} minute in the second half) were not included in this analysis to keep comparisons simple across matches, but could be of interest in further work. The distribution of total distance covered at the combined range of H_{ACC} and H_{DEC} from each group of outfield players in these six periods is illustrated in Figure 6.3.1. The highest mean for each position is in one of three 15-minute periods in the first half (P1–P3), while the lowest mean is always in the last 15-minute period (P6). The one-way ANOVA with Tukey-Kramer post-hoc test was performed to check whether there is a significant difference of the means between each pair of playing positions in the same 15-minute period, and the results are presented in Table 6.3.3. There is no statistical evidence that the central midfielders and the wide midfielders have different means of their total distance covered in all six 15-minute periods despite slightly higher values being observed from the central midfielders in most periods. Another pair with no significant difference in all periods is the forwards and the full-backs. The centre-backs, however, have the lowest means of the total distance covered in all these periods, and this difference is in general significantly different with other outfield players.

The means of the relative percentage of the total distance covered from all and each group of outfield players at H_{ACC} and H_{DEC} in four defined periods are presented by line plots in Figure 6.3.2. Both plots show very similar results such that their means in the peak period are approximately 40% higher than the overall mean, whereas the means in all other periods are close to the overall mean. During the 5_{POST}, players were likely to fully recover from temporary fatigue occurred in a short amount of time immediately after the peak period. Furthermore, there is no significant difference

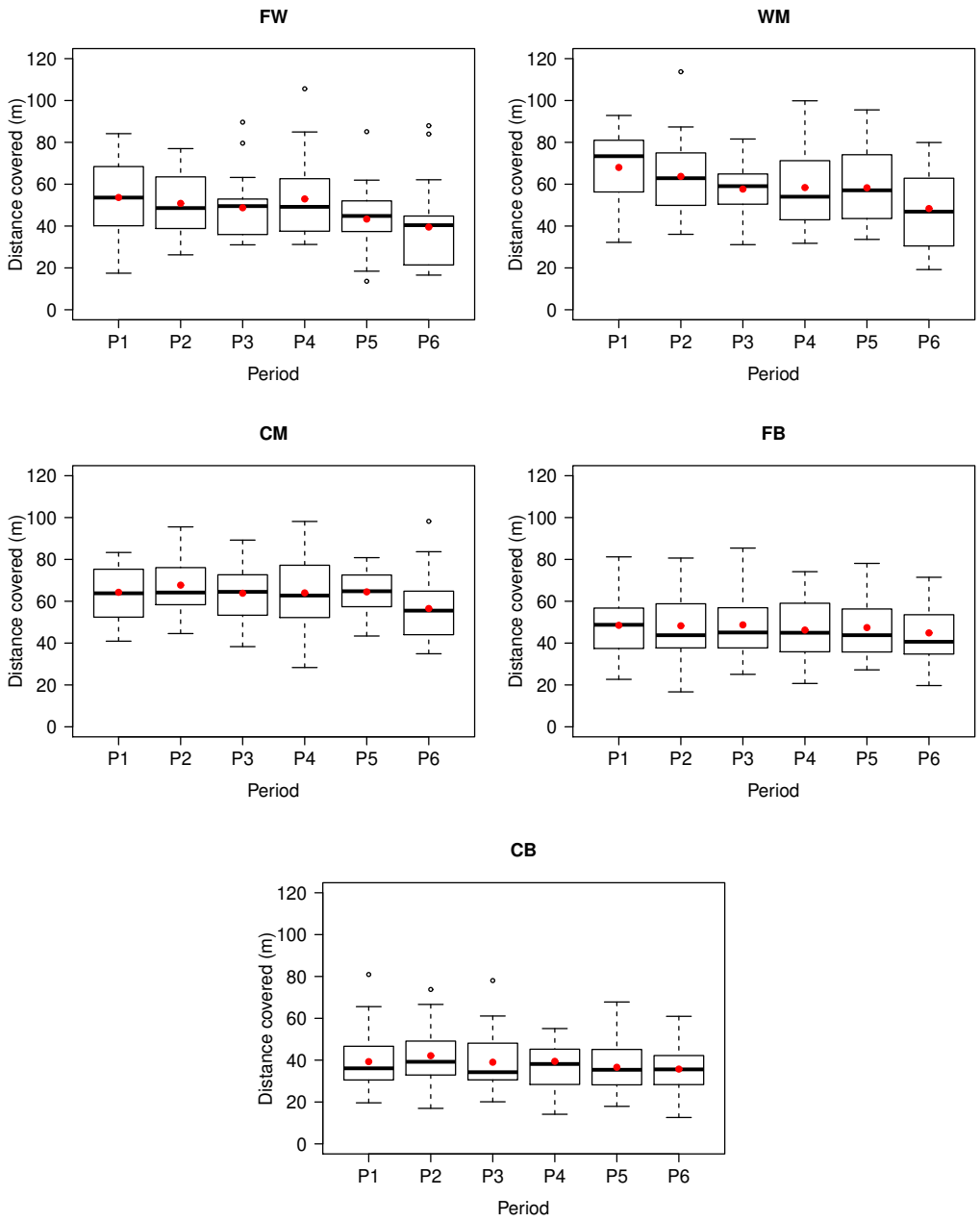

Figure 6.3.1: The box-and-Whisker plots of the total distance covered from all groups of outfield players at acceleration greater than 3 m/s^2 in magnitude in six 15-minute periods (P1–P6). A black thick line and a red dot in each box represent the median and the mean of the distance covered in that particular period, respectively.

Table 6.3.3: Results from the significant difference test of the means of the total distance covered at H_{ACC} and H_{DEC} between each pair of playing positions in six 15-minute periods (P1 to P6). "*" indicates significant difference, while "/" indicates no significant difference.

Position	FW	WM	CM	FB	CB
FW					
WM	**//*/				
CM	/**/**	//////			
FB	//////	**//*/	*****		
CB	*//*//	******	*****	//*/*/	

of the means between these defined periods excluding the peak period. This is in contrast to the findings by Akenhead et al. (2013), who found significant decreases of the mean before and after peak periods. One of the possible reasons for this is the different ways for calculating the overall mean of the distance between Akenhead et al. (2013) and our study, as given in Section 6.2.1. Akenhead et al. (2013) calculated this mean over the whole match, so we can expect that the mean from periods outside the peak period are significantly lower than the overall mean. On the other hand, we have removed the peak period from the calculation of the overall mean for a more meaningful measure. Even accounting for this difference, this only partly explained

the difference in results and the remaining differences could be simply due to different players and matches being analysed, or simply an insufficient data volume in both studies. We leave this open for future research and discussion, but we will move on to now explore significant turns to see if these have more effect on temporary fatigue than peak periods of acceleration and deceleration.

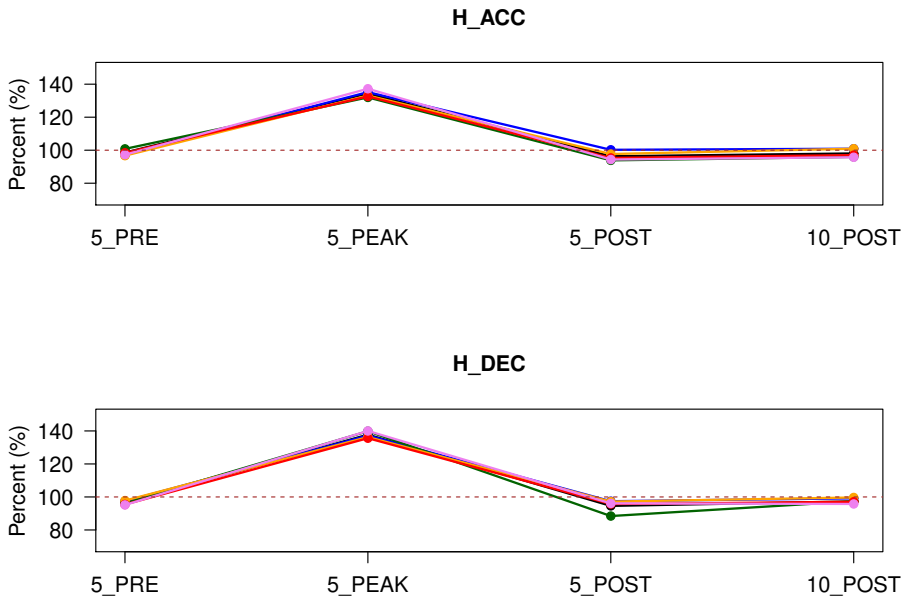

Figure 6.3.2: Line plots of the means of the relative percentage of the total distance covered in four defined periods from all and each group of outfield players at acceleration greater than 3 m/s^2 (upper panel) and less than -3 m/s^2 (lower panel). Both plots show the results from all outfield players (black), the forwards (blue), the wide midfielders (green), the central midfielders (orange), the full-backs (red), and the centre-backs (violet).

6.3.2 Significant Turns

Table 6.3.4 reports the mean and standard deviation of the ratio of the distance covered at high acceleration in magnitude, to the total distance, before and after significant turns from all and each group of outfield players (reported as percentages). The mean percentages decrease by 10% of their mean value in all cases, and this could be explained by the existence of fatigue or the avoidance of high energy output after significant turns. The wide and central midfielders have higher means than other groups of players in both periods before and after significant turns, while the centre-backs have the lowest means. This result is the same as what we derived from the means of the total distance covered in Table 6.3.2. However, our key new finding is that the mean before significant turns is significantly greater than after significant turns for each case of position as its p-value from the one-sided paired t-test is less than 0.05. Therefore significant turns appear to be a better predictor of temporary fatigue than peak periods of accelerations and deceleration as studied by Akenhead et al. (2013). This validates using significant turns as an interesting metric to monitor players' performance and fatigue.

Next, we calculated the summary statistics of this measure during significant turns (rather than before and after). The mean and standard deviation from all outfield players are 48.20% and 6.13%, respectively. Table 6.3.5 reports these summary statistics from each group of players, and their means are all around 45% to 55%, which are much higher than those before and after the significant turns. The one-way ANOVA with Tukey-Kramer post-hoc test was used to test the means between each pair of

Table 6.3.4: The mean and standard deviation (mean ± SD) in percentages of the ratio of the distance covered at high acceleration in magnitude, to the total distance, before and after significant turns, from all and each group of outfield players. P-values from the one-sided paired t-test of this measure before and after significant turns are also reported.

Position (number of players)	Before (%)	After (%)	P-value
All ($n = 149$)	4.25 ± 1.36	3.89 ± 1.36	< 0.001
FW ($n = 22$)	4.01 ± 1.40	3.60 ± 1.44	0.002
WM ($n = 24$)	4.89 ± 1.50	4.45 ± 1.53	< 0.001
CM ($n = 32$)	4.83 ± 0.96	4.42 ± 0.88	< 0.001
FB ($n = 35$)	4.12 ± 1.24	3.84 ± 1.28	0.004
CB ($n = 36$)	3.57 ± 1.32	3.28 ± 1.34	0.013

groups, however no significant difference was found in most pairs. This analysis confirms just how much energy is expended during periods of significant turns (over 10 times as much distance covered is under high acceleration in magnitude). The hypothesis is that this energy expenditure is responsible for the drop-off in the distance covered at high acceleration in magnitude seen after significant turns reported in Table 6.3.4—and this is something that can be studied further in future work.

Table 6.3.5: The mean and standard deviation (mean ± SD) in percentages of the ratio of the distance covered at high acceleration in magnitude to the total distance during significant turns from each group of outfield players.

Position	FW	WM	CM	FB	CB
During (%)	45.58 ± 8.13	50.99 ± 5.97	49.14 ± 3.46	48.86 ± 5.88	46.50 ± 6.07

Table 6.3.6 reports the total number of significant turns from players in different playing positions. The wide and central midfielders have higher mean than the others due to their playing roles in the field. The means between both halves are not significantly different in all cases of playing positions, according to the paired t-test.

Then we performed a regression analysis of the total number of significant turns in a professional match from players in different playing positions. We proposed three regression models in Section 6.2.2. The chi-square difference test with a significance level of 0.05 was used to test whether the more complex model with higher number of explanatory variables is significantly better to explain the expected number of significant turns than the simpler model with less number of explanatory variables. The test showed that models with the factor of playing positions (Models I and II) were significantly better than the model without this factor (Model III) (p-values < 0.001), however the multiplicative effects or interaction terms in Model I do not significantly improve the fit of the expected turns (p-value $= 0.450$). Thus, Model II is likely to be the best choice for modelling the expected number of significant turns.

Table 6.3.6: The mean and standard deviation (mean ± SD) of the total number of significant turns from all and each group of outfield players in both halves and the full match.

Position (number of players)	First half	Second half	Full match
All ($n = 149$)	11.19 ± 5.40	10.72 ± 6.09	21.91 ± 10.21
FW ($n = 22$)	8.14 ± 4.12	7.27 ± 4.93	15.41 ± 8.23
WM ($n = 24$)	12.54 ± 5.06	12.79 ± 6.09	25.33 ± 9.82
CM ($n = 32$)	17.44 ± 4.52	17.25 ± 5.44	34.69 ± 6.81
FB ($n = 35$)	9.43 ± 3.58	7.80 ± 3.96	17.23 ± 5.40
CB ($n = 36$)	8.31 ± 3.35	8.47 ± 3.68	16.78 ± 5.86

We used Model II to predict the expected number of significant turns from the distance covered and the groups of players. The Pearson correlation coefficient between these two explanatory variables is approximately 0.27, so that the multicollinearity does not exist in this model. The base level was set as the group of centre-backs (CB) due to their first alphabetical position. The model parameters were then estimated. We found that the model parameter of the distance covered is significantly different from zero. Furthermore, all groups except the forwards (FW) have no significant difference of their model parameters. These results were also reported when the base

level was set as other groups of outfield players. Thus, we could merge four groups of these players except the forwards into a single level of factor. A final generalised linear regression model of the expected number of significant turns was given by Model: $\log \lambda = \beta_0 + \beta_1 D + \beta_2 \mathbb{1}(\text{Group} = \text{“FW”})$, where all model parameters are significantly different from zero with p-values less than 0.05.

The estimated values and standard errors of all model parameters are presented in Table 6.3.7. The model can be expressed as $\log \lambda = -1.07 + (3.87 \times 10^{-4})D - (4.07 \times 10^{-1})\mathbb{1}(\text{Group} = \text{“FW”})$. Here we provide an example of how to interpret this model. If a player runs 10 kilometres in a full match, and he is a centre-back player, his expected number of significant turns is equal to $\exp(-1.07 + 3.87) = \exp(2.80) = 16.44$. This is also the expected number for all other players except the forwards with the same distance covered. For the forwards, their expected number of significant turns is equal to $\exp(-1.07 + 3.87 - 0.407) = \exp(2.393) = 10.95$. One possible reason for the difference of the expected number of significant turns between forwards and all other outfield players is the team's playing style, especially if focusing on more defending than attacking strategies in most matches. The graphical representation of the quasi-Poisson regression model for our significant turns data is illustrated in Figure 6.3.3. The regression line of the expected number of significant turns for the forwards and its 95% confidence interval are clearly lower than those for other groups of outfield players when the distance covered is between 8 and 14 kilometres, which is the normal observed range for all outfield players playing in the full match. This model therefore provides a simple way of linking expected number of significant turns with player position and total distance covered. Further analyses could explore the

impact this relationship (and deviations that players show away from this relationship) might have on performance and injury. For example, it could be investigated whether a higher than expected number of significant turns is a contributing factor to player injury or long-term fatigue across a season.

Table 6.3.7: The estimated values and standard errors of model parameters of the generalised linear model with the quasi-Poisson regression (Model II) for the prediction of the expected number of significant turns.

Model parameter	Estimated value	Standard error
β_0 (Constant)	-1.07	2.84×10^{-1}
β_1 (Distance covered)	3.87×10^{-4}	2.56×10^{-5}
β_2 (Position: FW)	-4.07×10^{-1}	7.47×10^{-2}

6.3.3 Nonparametric Measures

Table 6.3.8 reports the summaries of IV values from the acceleration time series data from all and each group of outfield players in both halves and the full match. Most acceleration data have their IV values close to two, meaning that they are likely to behave like Gaussian white noise. However, the mean of IV values from the forwards is significantly lower than other groups of players in the full match. So, the forwards are likely to have less fluctuation of their acceleration time series data in every second

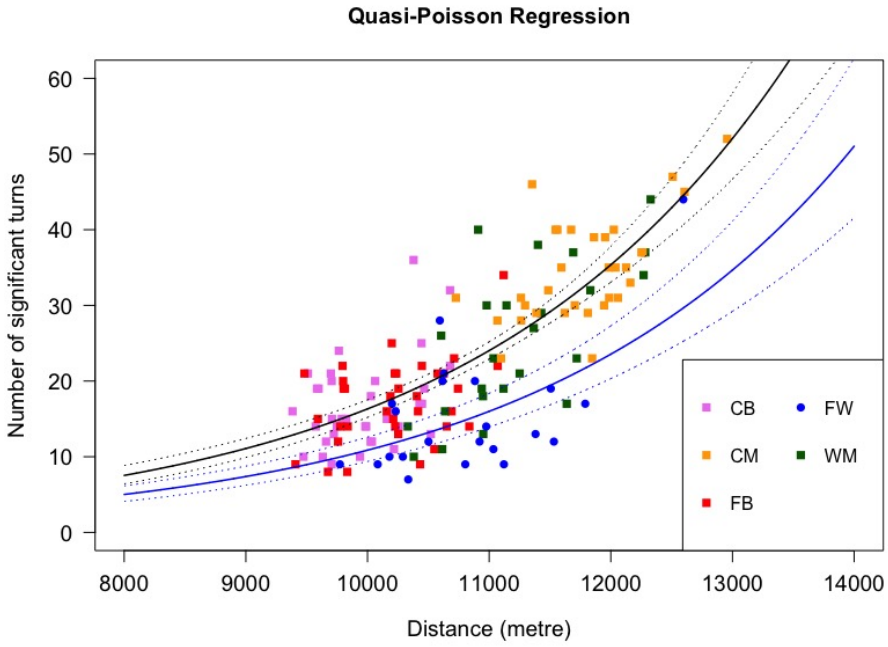

Figure 6.3.3: The generalised linear model with the quasi-Poisson regression of the expected number of significant turns (y-axis) with respect to the distance covered in the full match (x-axis) and the groups of players. The data from forwards are presented with blue circles, and their regression line along with its 95% confidence interval ($\pm 1.96 \times$standard error) are presented with solid and dotted blue lines, respectively. The data from other groups of outfield players are presented with different colours of squares depending on their playing positions, and their regression line (as one joint group) along with its 95% confidence interval are presented with solid and dotted black lines, respectively.

than other players. The central midfielders, on the other hand, have significantly higher IV values than all other playing positions, and hence they are likely to have larger fluctuation of their data per second. We also observed that IV values from acceleration data are positively correlated with the total number of significant turns

in the match, and this is reasonable as these turns are triggered from the rapid change of acceleration over a short period of time. In addition to the comparison of IV values between groups of players, we performed the paired t-test with the significance level of 0.05 to check whether there is a difference of the means of IV values between both halves. Although the means in the first half are higher than the second half, there is no significant difference between them in all groups except the centre-backs (with p-value = 0.045).

Table 6.3.8: The mean and standard deviation of IV values from the acceleration time series data from all and each group of outfield players during the first half, the second half, and the full match excluding the additional times.

Position (number of players)	First half	Second half	Full match
All ($n = 149$)	2.05 ± 0.09	2.02 ± 0.09	2.04 ± 0.08
FW ($n = 22$)	1.95 ± 0.08	1.93 ± 0.07	1.94 ± 0.07
WM ($n = 24$)	2.05 ± 0.06	2.01 ± 0.11	2.04 ± 0.08
CM ($n = 32$)	2.12 ± 0.05	2.10 ± 0.07	2.11 ± 0.05
FB ($n = 35$)	2.01 ± 0.08	1.98 ± 0.07	2.00 ± 0.06
CB ($n = 36$)	2.07 ± 0.07	2.04 ± 0.06	2.06 ± 0.05

Table 6.3.9 reports the summaries of the DFA scaling exponents (α) from the ac-

celeration time series data from all and each group of outfield players in both halves and the full match. The forwards have the highest mean of α, while the central midfielders have the lowest mean of α in the full match. No significant difference of the means is found for all outfield players between the first half and the second half. Because $\alpha < 1$, all acceleration data from football players are stationary according to this analysis, meaning that their statistical properties do not change over time. In addition, we observe that $\alpha > 0.5$ meaning there is some evidence of long memory in the time series. Several normality tests including the graphical representations of histogram and quantile-quantile (Q-Q) plot, and the Shapiro-Wilk test were applied to all data, however they provided clear evidence of non-normally distributed acceleration data in most cases. This explains why we see some inconsistency between Table 6.3.8 (where IV values of 2 are consistent with white noise) and Table 6.3.9 (where here we see evidence of persistence and long memory as $\alpha > 0.5$ which is consistent with $H > 0.5$). What is consistent between the results however is the fractional behaviour, where a lower α is consistent with a higher IV for time series that are more spikey and volatile (and vice versa for smoother time series). Therefore again we observe that the central midfielders have the most fragmented and spikey accelerations, and the forwards the least. Therefore we have found utility in the non-parametric methods used traditionally for accelerometry data in a medical setting to also be used for sports tracking data such as in this chapter - in particular IV and DFA reveal interesting structures that separate player behaviour, and we believe this warrants further investigation and usage in future.

Table 6.3.9: The mean and standard deviation of the scaling exponents (α) from the acceleration time series data from all and each group of outfield players during the first half, the second half, and the full match excluding the additional times.

Position (number of players)	First half	Second half	Full match
All ($n = 149$)	0.66 ± 0.03	0.67 ± 0.04	0.66 ± 0.03
FW ($n = 22$)	0.69 ± 0.03	0.70 ± 0.03	0.69 ± 0.02
WM ($n = 24$)	0.66 ± 0.03	0.68 ± 0.04	0.67 ± 0.03
CM ($n = 32$)	0.63 ± 0.02	0.65 ± 0.02	0.64 ± 0.02
FB ($n = 35$)	0.67 ± 0.04	0.68 ± 0.03	0.68 ± 0.02
CB ($n = 36$)	0.64 ± 0.03	0.66 ± 0.02	0.65 ± 0.02

6.3.4 Comparison of Nonparametric Measures

Since both IV and DFA scaling exponent (α) are used to measure the degree of fluctuation of time series data, it is expected that their estimated values are strongly correlated, as we just discussed. In addition to these, the total number of significant turns is considered due to its positive correlation with IV. Table 6.3.10 reports the Pearson correlation coefficients (ρ) from each pair of all these variables derived from the full match. The results from both halves are close to the full match, so we will

not report them here. According to this table, IV is strongly negatively correlated with α for all cases of position, and such linear relationship agrees with the finding from our previous work on the acceleration data from patients with and without advanced dementia (Suibkitwanchai et al., 2020). The correlation coefficients between the total number of significant turns and both nonparametric measures of the degree of fluctuation also represent moderate to strong correlation, where stronger values are found from the relationship with IV. It is interesting that the wide midfielders have the strongest correlation coefficients, while the full-backs have the weakest correlation coefficients for both pairs of the total number of significant turns and nonparametric measures. Figure 6.3.4 shows how all three variables including IV, DFA scaling exponent (α), and the total number of significant turns in a full match are related by using the scatter plots with different colour of points depending on playing positions. It is clear that both nonparametric measures from IV and DFA have very strong negative linear relationship. The forwards (blue circle dots) are likely to have low IV but high DFA scaling exponents, while the central midfielders (orange square dots) are likely to have high IV but low DFA scaling exponents, as already discussed. The measures from other playing positions look inseparable although their points approximately lie on a line with negative slope. The points between the total number of significant turns and both measures of the degree of fluctuation are more dispersed with moderate to strong linear relationship.

Table 6.3.10: The Pearson correlation coefficients (ρ) for all and each group of outfield players from each pair of three different variables: IV, DFA scaling exponent (α), and the total number of significant turns in a full match.

ρ	All	FW	WM	CM	FB	CB
(IV, α)	-0.96	-0.96	-0.94	-0.89	-0.95	-0.93
(IV, turns)	0.72	0.66	0.90	0.62	0.51	0.60
(α, turns)	-0.64	-0.62	-0.81	-0.55	-0.43	-0.59

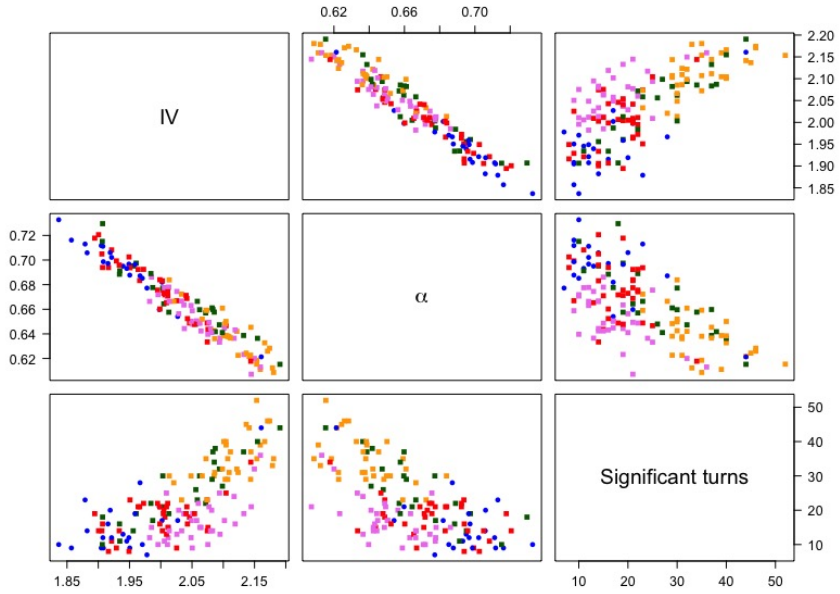

Figure 6.3.4: The scatter plots showing the relationship of measures from each pair of three variables: IV, DFA scaling exponent (α), and the total number of significant turns in a full match. The different colours of points represent the different groups of outfield players including the forwards (blue), the wide midfielders (green), the central midfielders (orange), the full-backs (red), and the centre-backs (violet). The shapes of points for different players correspond to those in the regression plot of Figure 6.3.3.

Chapter 7

Conclusions

We summarise all contributions that have been discussed in this thesis. Future work and several suggestions of further research are also made.

7.1 Contributions

This thesis consists of several works regarding the estimation of parameters of models for time series with long memory, and nonparametric statistical measures related to long memory and fluctuations from high-frequency data of two applications. Our parametric models are different types of fractional processes including FGN, FBM, and the FD process. These processes exhibit their long-memory behaviour for different ranges of their controlling parameters, and the key controlling parameter is (or can be related to) the Hurst exponent, as we detailed in Chapter 2.

In Chapter 3, we provided several parametric estimation methods including the WLE and the DWLE. This debiased method was proposed by Sykulski et al. (2019)

to reduce the bias of estimates while maintaining the same rate of consistency, and the same computational cost, as the regular Whittle likelihood. They applied this estimator to Gaussian and non-Gaussian processes with short-memory time series. In this thesis, however, we explored the use of such an estimator to both short and long-memory time series. We explicitly detailed how these estimators should be implemented for processes such as FGN, FBM and the FD process.

In Chapter 4, we performed a detailed simulation analysis and estimated the Hurst exponent along with the sample variance of FGN, which is a stationary process. We found that the DWLE provides very close parameter estimates to the WLE. However, if this process exhibits long-memory behaviour with the Hurst exponent greater than 0.5, the measurement errors including the MAE and the RMSD of estimates from the DWLE are the lowest among all estimation methods considered. In addition, another advantage of this estimator is its requirement of less average computational time than the WLE, which has much higher time especially when the approximated aliased spectrum is used. The DWLE, on the other hand, bypasses the need to use the power spectral density function by using the autocovariance instead to fit the process, which is especially advantageous in the case of FGN. The estimates from the DWLE are likely to be $\sqrt{n}$-consistent (or converge at a rate of $1/\sqrt{n}$) according to the linear regression plot from simulated FGNs with selected true values of the Hurst exponent and lengths n. However, this convergence rate from the simulation holds for any value of the Hurst exponent between zero and one. The approximation of the standard error of estimates, and the use of the local Hurst exponent to describe the change of the fractional behaviour over time of real-world financial data, were also

explored in this chapter.

We also studied the estimation from other fractional processes in Chapter 4. First, we tried to model FGN with the FD process. However, the relationship between the estimates of long-memory parameters from both processes does not agree with its theoretical concept, where the difference is exactly equal to 0.5. The misspecification of the model seems to have an effect on the estimation of these parameters, with the misspecification effect most exacerbated as the Hurst parameter goes towards its extremes of zero or one. We also compared the estimates derived from directly fitting FBM to the data, to those from fitting FGN to the first-order difference of such data. The DWLE with tapering provides the least bias among all estimation methods from directly fitting FBM, which is similar to the result from fitting FGN to the differenced data. However, its very high cost for the computation of the expected periodogram makes the direct method less practical than first-order differencing and fitting FGN. Thus, we learnt from this chapter that (1) it is better to estimate long-memory parameters from FGN than FBM when considering both measurement errors and computational time, (2) the DWLE is effective in reducing the bias of long-memory parameters, (3) model misspecification can make for very poor estimation of long-memory parameters, and (4) the tapering technique can improve the accuracy of estimates especially for a process with high dynamic range of its spectrum.

Next, we performed some detailed application analyses in Chapters 5 and 6. Parametric models were not feasible to fit the structures of data from both chapters, so instead we focused on nonparametric measures including those that captured long-memory and fractional behaviours in time series.

In Chapter 5, we focused on the time series of accelerometry data from patients with advanced dementia. We refer the reader to Section 5.3 for a detailed conclusion. In summary, the key finding was that fractional and long-memory properties of the time series of individuals could effectively be separated between those with and without advanced dementia by several nonparametric measures. These included the use of IV and the scaling exponent from DFA as nonparametric summary statistics. We devoted some time in this chapter to study the optimal ways in which IV should be subsampled to optimally tune this statistic in practice. In addition, we also looked at the daily "circadian" rhythm of the time series, and explored two other measures, IS and PoV, to quantify the strength of this cycle. Both measures could also be used to separate individuals with and without advanced dementia.

In Chapter 6, we analysed the high-frequency tracking data from footballers during professional matches. We followed the analysis from Akenhead et al. (2013) by studying the distance covered at different ranges of speed and acceleration, and in different periods of the match. Due to temporary fatigue, the substantial decrease of the distance covered at high acceleration in magnitude during the 5-minute period after the peak period was observed from both Akenhead et al. (2013) and our analysis. However, we improved this measure and showed that in fact there is no significant difference of such measure between before and after peak periods, unlike Akenhead et al. (2013).

In Chapter 6, we also introduced a new measure of fatigue by comparing the rate of the distance covered at high acceleration in magnitude before and after "significant turns", which are events of sudden change from high deceleration to high acceleration

in a short amount of time. The application of this measure from significant turns is likely to outperform that from the peak period of distance covered in Akenhead et al. (2013), in terms of identifying periods of fatigue. This is due to evidence of significant decreases of the measure from before to after significant turns for every playing position of outfield player. Another analysis from significant turns was to fit their count data with the quasi-Poisson regression model. All outfield players except the forwards shared the same model, which predicted the number of significant turns from their distance covered in the full match. Lastly, we implemented two nonparametric measures including IV and the scaling exponent from DFA on acceleration data of players. A comparison of these results indicated that the data is unlikely to be Gaussian due to inconsistencies in values obtained, thus justifying not implementing Gaussian parametric models like FGN and instead focusing on nonparametric measures. However, we found a strong negative correlation between these two measures like Suibkitwanchai et al. (2020), and several playing positions of players were separable from the use of these measures.

7.2 Future Works

7.2.1 Theoretic Properties of the Debiased Whittle Likelihood Estimator for Long-Memory Processes

In Figure 4.2.5, we showed the linear relationship between standard deviation of the estimated Hurst exponent and the length of time series of FGN in log-log space. We

suggested that the estimate of the Hurst exponent with the DWLE is $\sqrt{n}$-consistent according to the slope of linear regression lines from this figure, and these lines represent the results from simulated FGN with $H_T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. In Section 4.2 of this figure, we also stated that we currently have no completed mathematical proof for this consistency of estimates of fractional process with long-memory behaviour. However, we have done some parts of this proof by following the proof for short-memory time series in the Supplementary Material of Sykulski et al. (2019). This Supplementary Material consists of several theorems and lemmas required to prove consistency and establish convergence rates. Some assumptions made for these theorems and lemmas are not satisfied by time series with long-memory, for example, an assumption is made that the power spectral density function is bounded both above and below some constants for all ranges of frequencies less than or equal to the Nyquist frequency. For long memory, the power spectral density function is unbounded at the zero frequency (see Equation (2.3.6)), therefore we cannot make immediate use of the theory developed in Sykulski et al. (2019) for this case. We have adjusted some of these theorems and lemmas to fit with time series with long-memory behaviour. One of these is extending Lemma 8 in the Supplementary Material of Sykulski et al. (2019) about the variance of linear combinations of values of the periodogram, $I(\omega)$, but this time extended to a discrete-time FGN given by $X = \{X_t\}_{t=1}^{n}$. Since the spectrum of long-memory time series approaches infinity at the zero frequency, the upper boundary of spectrum is not defined. Thus, we seek a new approach for the

upper boundary of this variance, and this is given by

$$\mathrm{Var}\left[\frac{1}{n}\sum_{\omega\in\mathbf{\Omega}} a_n(\omega)I(\omega)\right] \leq \mathrm{Var}\left[\frac{a_{max}}{n}\sum_{\omega\in\mathbf{\Omega}} I(\omega)\right] \leq \frac{a_{max}^2}{n^2}\mathrm{Var}\left[\sum_{t=1}^{n} X_t^2\right], \qquad (7.2.1)$$

where $\mathbf{\Omega}$ is a finite set of discrete Fourier frequencies, i.e., $\mathbf{\Omega} \equiv \{\omega_j\}_{j=1}^n$ for each $\omega_j = 2\pi(j - \lceil n/2 \rceil)/n$, and a_{max} is the upper bound of $a_n(\omega)$ for all $\omega \in \mathbf{\Omega}$, which is a deterministic function.

We seek to expand this equation, to express it in terms of the autocovariance sequence or s_τ. We will use Isserlis' Theorem and bound the behaviour of s_τ by substituting $s_\tau \leq k\tau^{2(H-1)}$, where k is a constant and τ is a positive integer. This condition of the autocovariance sequence is derived from its asymptotic property from Equation (2.6.3). The summation in the final term of Equation (7.2.1) is expanded as follows:

$$\begin{aligned}
\frac{a_{max}^2}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i^2\right] &= C\sum_{j=1}^{n}\sum_{k=1}^{n}\mathrm{Cov}[X_j^2, X_k^2] \\
&= C\Big\{n\mathrm{Var}[X_t^2] + 2(n-1)\mathrm{Var}[X_t^2, X_{t-1}^2] + \dots\Big\} \\
&= C\Big\{n\big(\mathbb{E}\left[X_t^4\right] - \mathbb{E}\left[X_t^2\right]\mathbb{E}\left[X_t^2\right]\big) \\
&\quad + 2(n-1)\big(\mathbb{E}\left[X_t^2X_{t-1}^2\right] - \mathbb{E}\left[X_t^2\right]\mathbb{E}\left[X_{t-1}^2\right]\big) + \dots\Big\} \\
&= C\Big\{2n\mathbb{E}\left[X_t^2\right]^2 + 4(n-1)\mathbb{E}\left[X_tX_{t-1}\right]^2 + \dots\Big\} \\
&= C\Big\{2ns_0^2 + 4(n-1)s_1^2 + 4(n-2)s_2^2 + \dots + 4s_{n-1}^2\Big\} \\
&\leq C\Big\{4n\sum_{\tau=0}^{n-1} s_\tau^2\Big\} \\
&\leq 4nC\Big\{s_0^2 + k^2\sum_{\tau=1}^{n-1}\frac{1}{\tau^{4(1-H)}}\Big\}, \qquad (7.2.2)
\end{aligned}$$

where $C = a_{max}^2/n^2$. For very large n, the above series converges when $4(1-H)$

is greater than one, i.e., $H < 0.75$. Hence, we can show that the variance of linear combinations of values of the periodogram in Equation (7.2.1) scales as $\mathcal{O}(1/n)$ when $0 < H < 0.75$. Although the p-series diverges when $0.75 \leq H < 1$, we can approximate it and show that the variance at large n instead scales as $\mathcal{O}(1/n^{4(1-H)})$ when $0.75 \leq H < 1$ (See Appendix C).

This result suggests a different approach should be applied to long-memory time series of FGN with $0.5 < H < 0.75$, and $0.75 \leq H < 1$ for proving the consistency and the rate of convergence. The different establishment of the central limit theorem between these two ranges of H was also mentioned in some research papers such as Yajima (1985) and Lobato and Robinson (1996). We have adjusted several other small lemmas to fit with time series with long-memory behaviour, however most of them are similar to what were explained for short-memory time series in the Supplementary Material of Sykulski et al. (2019). Overall, we leave the final complete proof of consistency and the convergence rate to future work. Although Sykulski et al. (2019) provided different big O expressions for Lemma 8, the convergence rate of parameter estimates should be mathematically the same which is $\sqrt{n}$-consistent. This has been evidenced from our results from simulated FGNs, and is also consistent with the established convergence rates of the WLE for long-memory time series (Fox and Taqqu, 1986).

7.2.2 Methods and Models for Long-Memory Processes

In this thesis, we provided several methods for the estimation of parameters of fractional processes which can exhibit either short or long-memory behaviour. There

exist other estimators apart from those in Chapter 3 that could be used to estimate long-memory parameters. Several examples of these are nonparametric estimators such as the aggregated variance estimator (Beran, 1994), the absolute moments estimator (Taqqu et al., 1995), and the wavelet estimator (Bardet et al., 2000). However, if a model is correctly specified, these nonparametric estimators will not be the best choice for fractional processes since there is a high possibility that their bias and variance are greater than those from parametric estimators provided in Chapter 3. On the other hand, non-parametric estimates will in general be more robust to misspecified models, and may be more robust to extremes, outliers, and non-Gaussian or non-stationary behaviour. Comparing estimation results from nonparametric estimators will be conducted in future work, as well as comparisons of robustness against parametric methods.

We fitted several fractional processes including FGN, FBM, and the FD process for modelling a finite time series, and estimated their long-memory parameters in this thesis. However, there are several other models that can be used to model time series with long memory. Because the FD process is equivalent to the ARFIMA$(0, d, 0)$ process, it will often be the case that the generalisation of this process, the ARFIMA(p, d, q) process, can explain long-memory behaviour of time series with some specific values of p, d, and q. Another process is the fractionally integrated generalised autoregressive conditionally heteroskedastic (FIGARCH) process, which was proposed by Baillie et al. (1996). This process is often used as a long-memory model for volatility in finance. It would be of interest to see how our estimation methods can be applied to the generalised ARFIMA(p, d, q) process and the FIGARCH process, and how accurate

the estimates of their parameters are obtained. Such analyses will be performed in future work.

7.2.3 Stationarity vs. Nonstationarity

In this thesis, we estimated the Hurst exponent of FBM and financial data by either directly fitting FBM to the time series data or fitting FGN to its first-order difference. This means that we assumed such data to be nonstationary with fractional behaviour, where simulated FBMs were generated by the simulation procedure explained in Section 4.1. However, for any given real-world time series, we do not know whether or not it is stationary, despite the fact that most real-world time series will usually have some nonstationary behaviour present. The stationarity of time series data can be checked by several methods. One possible way is to apply DFA to this time series and calculate its scaling exponent. In Chapter 5, we provided the details that this scaling exponent is between 0 and 1 for a stationary time series, whereas it is between 1 and 2 for a nonstationary time series. Another method is to use the slope from the LPRE. In Chapter 3, we explained that this slope is likely to be between -1 and 1 for a stationary time series, and between -3 and -1 for a nonstationary time series. However, for $1/f$ noise, the slope is -1 and is hence on the boundary, and therefore there is a possibility that both the scaling exponent and the LPRE can misclassify the stationarity property of time series. We have tried possible solutions including (1) the comparison of log-likelihood values (see Equation (3.2.1) for the WLE, and Equation (3.3.1) for the DWLE) between fitting FGN with H close to one and FBM with H close to zero, and (2) the use of the augmented Dickey–Fuller (ADF) test (the adap-

tation from a test in Dickey and Fuller (1979), which was originally used for checking the stationarity of AR processes). However, both solutions cannot fully resolve the problem especially for $1/f$ noise with nonstationary time series. This "boundary" issue at the limit between stationarity and nonstationarity has been studied elsewhere in the time series community in the context of ARFIMA models when d is close to 0.5 (Griffin et al., 2011). We will further investigate this problem in future work.

7.2.4 Future Works on Fatigue Analysis

In Chapter 6, we studied the impact of fatigue on footballers' performance by calculating some measures including the distance covered at high acceleration in magnitude, and the total number of significant turns from different groups of players. Because the most important outcome in a football match is the final result such as win, draw, or lose, we may use our measures on fatigue analysis to try to predict this result. A possible model for this prediction is the multinomial logistic regression calculating the probability of each outcome, which is a categorical dependent variable, from a set of covariates such as the total number of significant turns and the distance covered at high acceleration in magnitude of all outfield players in the full match. We may also consider these measures between the first half and the second half or between groups of players as different covariates.

All our data were from home games. However, the company we worked with may commence recording these data from away games in the future, or from other teams and stadia. If this is the case, we may possibly compare our measures between home and away games, and between various teams and types of matches, and this will be

left for further analysis in our future works once such data becomes available.

Appendix A

Supplementary for Chapter 5

A.1 Demographic details of participants

Table A.1.1: Demographic information of all participants from Froggatt et al. (2018). The 26 participants from which we have valid accelerometry data are from this study.

Characteristic	Non-intervention	Intervention
Sex, n (%)		
Male	9 (64.3)	8 (44.4)
Female	5 (35.7)	10 (55.6)
Age (years), mean±SD	84.7 ± 6.4	79.0 ± 10.5
Marital status, n (%)		
Married	11 (78.6)	7 (38.9)
Single	0	3 (16.7)
Widowed	3 (21.4)	8 (44.4)
Ethnicity, n (%)		
Black African Caribbean	0	1 (5.6)
White	14 (100.0)	17 (94.4)
Dementia diagnosis, n (%)		
Alzheimer's disease	7 (50.0)	7 (38.9)
Vascular dementia	1 (7.1)	5 (27.8)
Dementia with Lewy bodies	2 (14.3)	0
Unspecified dementia	4 (28.6)	6 (33.3)
Length of stay (years), mean±SD	2.9 ± 2.6	2.1 ± 1.4

A.2 Nonparametric results

A.2.1 Summary statistics

Table A.2.1: Mean, standard deviation (SD), minimum and maximum values of all statistical measures used in Chapter 5 for different groups of participants. The dementia group refers to the combined non-intervention and intervention groups.

Statistical measures	Groups	Mean	SD	Min.	Max.
IS	Non-intervention	0.029	0.019	0.003	0.063
	Intervention	0.039	0.025	0.005	0.082
	Dementia	0.035	0.023	0.003	0.082
	Without Dementia	0.147	0.054	0.062	0.257
IV	Non-intervention	1.605	0.137	1.424	1.856
	Intervention	1.519	0.186	1.036	1.775
	Dementia	1.555	0.169	1.036	1.856
	Without Dementia	1.028	0.301	0.376	1.487
α (overall)	Non-intervention	0.816	0.057	0.722	0.946
	Intervention	0.878	0.050	0.808	0.965
	Dementia	0.852	0.061	0.722	0.965
	Without Dementia	1.005	0.098	0.843	1.174
α (daytime)	Non-intervention	0.820	0.066	0.717	0.983
	Intervention	0.875	0.053	0.792	0.962
	Dementia	0.852	0.064	0.717	0.983
	Without Dementia	1.006	0.097	0.842	1.175
α (nighttime)	Non-intervention	0.806	0.067	0.710	0.891
	Intervention	0.897	0.053	0.825	1.004
	Dementia	0.859	0.074	0.710	1.004
	Without Dementia	0.905	0.083	0.741	1.043
$\text{PoV}^{(F)}$	Non-intervention	0.027	0.031	0.001	0.101
	Intervention	0.049	0.046	0.002	0.138
	Dementia	0.040	0.041	0.001	0.138
	Without Dementia	0.152	0.065	0.077	0.305
$\text{PoV}^{(H)}$	Non-intervention	0.055	0.036	0.007	0.122
	Intervention	0.078	0.052	0.012	0.167
	Dementia	0.068	0.047	0.007	0.167
	Without Dementia	0.224	0.071	0.108	0.338

A.2.2 Correlation between statistical measures

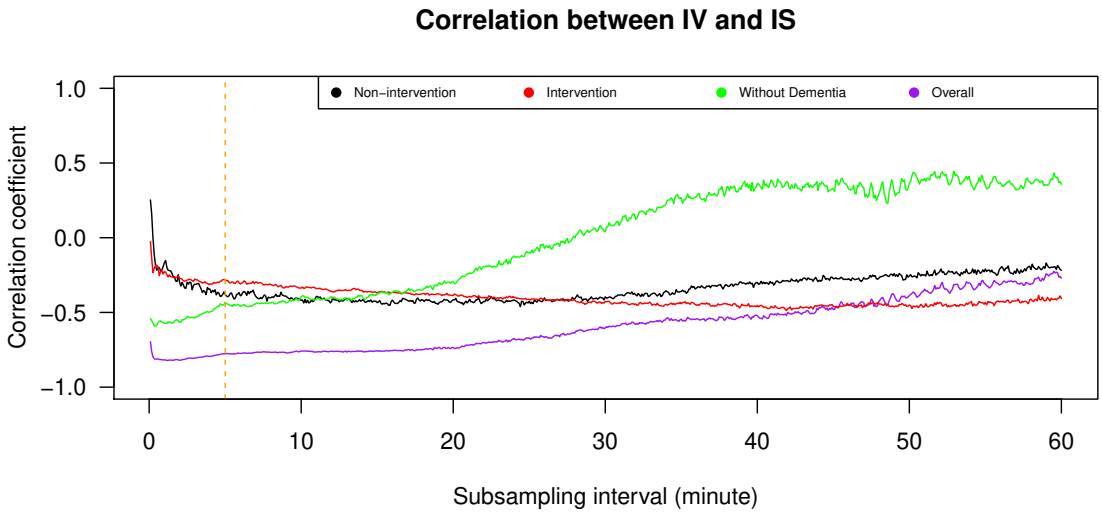

Figure A.2.1: Line plots showing Pearson's correlation coefficients between IV and IS for the ENMO data with subsampling interval for calculating IV varied from 5 seconds to 60 minutes. The non-intervention group, the intervention group, the group of individuals without dementia, and the overall group are presented by black, red, green, and purple lines, respectively. The orange dashed line refers to 5 minutes and indicates our recommended subsampling interval for calculating IV.

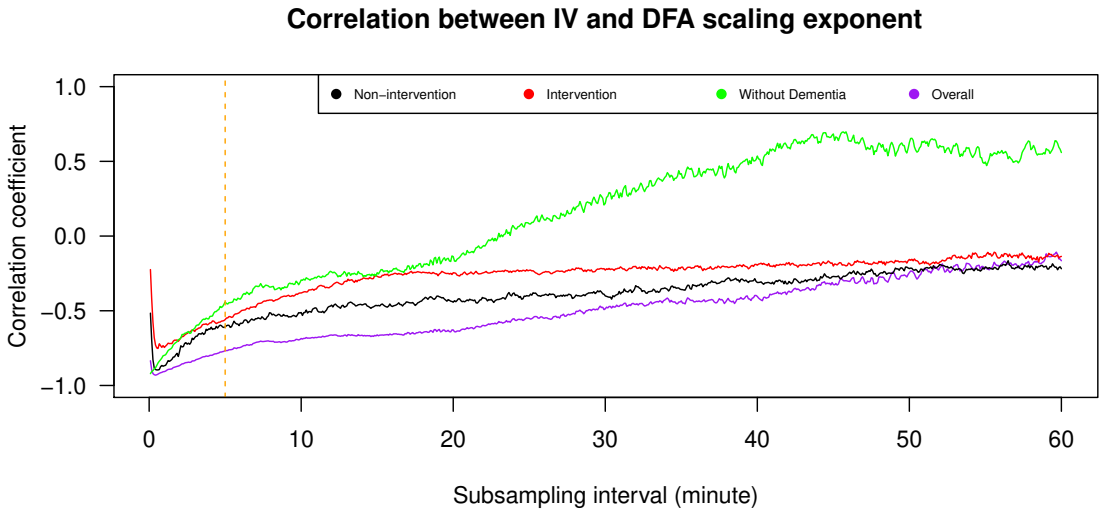

Figure A.2.2: Line plots showing Pearson's correlation coefficients between IV and the DFA scaling exponent for the ENMO data with subsampling interval for calculating IV varied from 5 seconds to 60 minutes. The non-intervention group, the intervention group, the group of individuals without dementia, and the overall group are presented by black, red, green, and purple lines, respectively. The orange dashed line refers to 5 minutes and indicates our recommended subsampling interval for calculating IV.

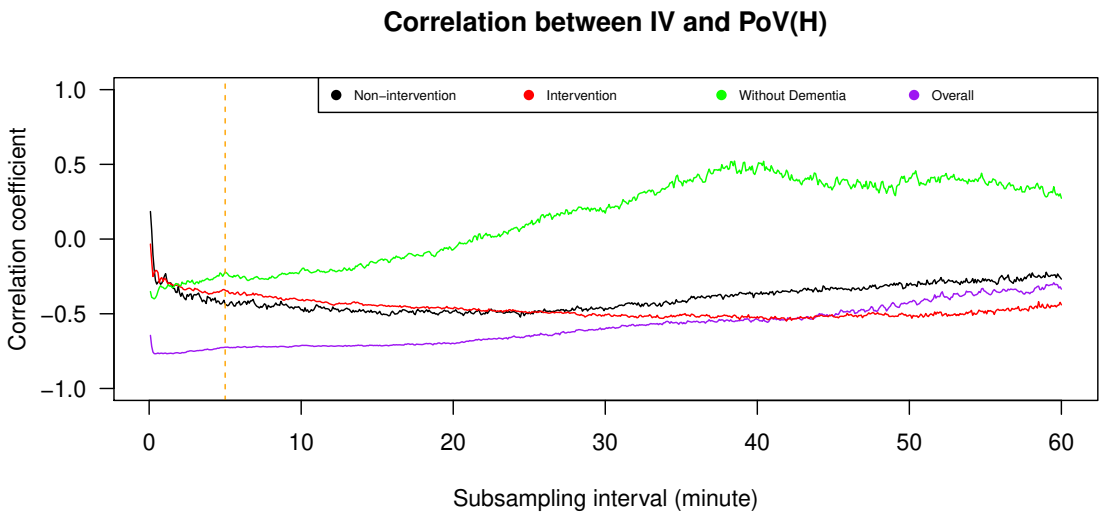

Figure A.2.3: Line plots showing Pearson's correlation coefficients between IV and PoV around the first four harmonic frequencies for the ENMO data with subsampling interval for calculating IV varied from 5 seconds to 60 minutes. The non-intervention group, the intervention group, the group of individuals without dementia, and the overall group are presented by black, red, green, and purple lines, respectively. The orange dashed line refers to 5 minutes and indicates our recommended subsampling interval for calculating IV.

Appendix B

Supplementary for Chapter 6

B.1 Ratio of the distance covered at high acceleration in magnitude v.s. Time period

The upper panel of Figure B.1.1 illustrates the mean of the ratio of the distance covered at high acceleration in magnitude to the total distance in one second to two minutes after significant turns from all and each group of players. Not much difference between all lines from these players is observed. There is a large decrease in the first 20 seconds and an almost constant relationship after that. The difference between the means before and after significant turns is also visualised in the lower panel of Figure B.1.1. The mean after is lower than before significant turns in the first few seconds, but they are approximately equal later. The selection of one minute as the time duration for our statistical analysis of this measure was appropriate due to its stability at one minute as shown in the figure.

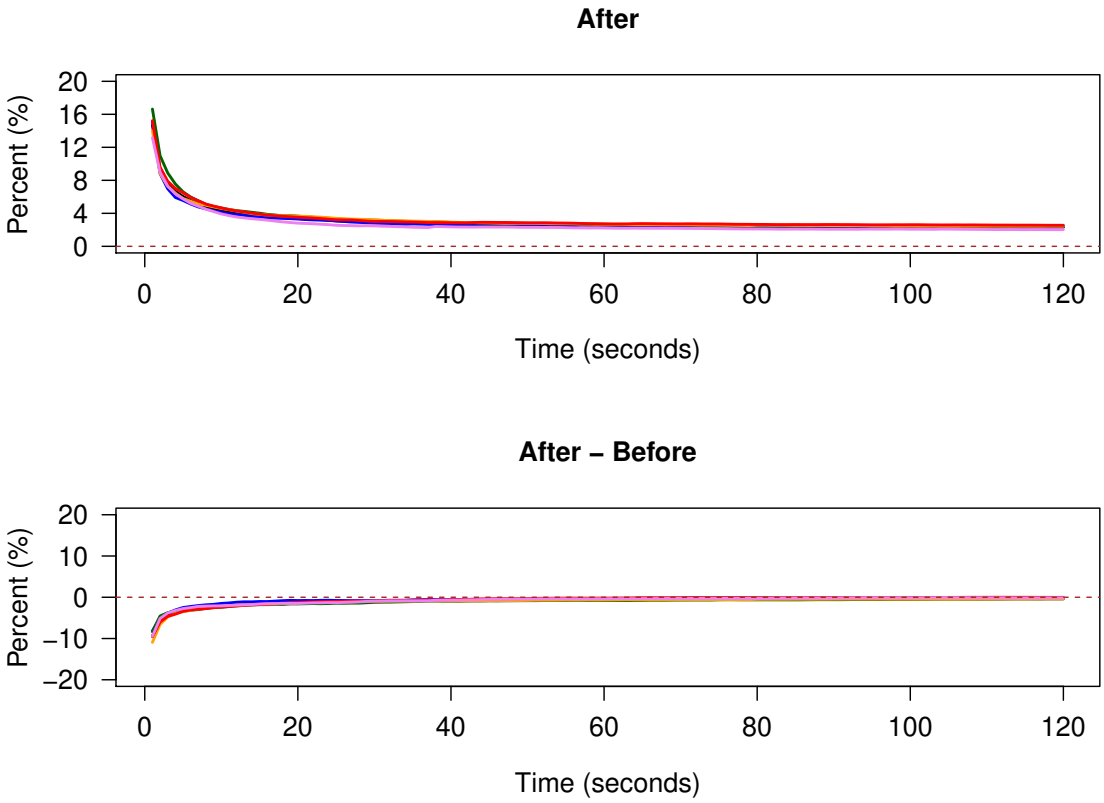

Figure B.1.1: The upper panel shows line plots of the means in percentages of the ratio of the distance covered at high acceleration in magnitude to the total distance in one second to two minutes after significant turns, while the lower panel shows line plots of the difference between the means before and after significant turns in one second to two minutes. These line plots include the forwards (blue), the wide midfielders (green), the central midfielders (orange), the full-backs (red), the centre-backs (violet), and all these players (black). A brown dashed line is used as the zero-reference line.

Appendix C

Supplementary for Chapter 7

C.1 Time complexity of the variance of linear combinations of periodogram for FGN with a high value of the Hurst exponent

The p-series in Equation (7.2.2) diverges when $0.75 \leq H < 1$. However, for very large n, this can be approximated by an asymptotic expansion with the use of Euler-Maclaurin formula as

$$\sum_{\tau=1}^{n-1} \frac{1}{\tau^p} \approx \zeta(p) - \frac{1}{(p-1)(n-1)^{p-1}} + \frac{1}{2(n-1)^p} - \sum_{i=1}^{\infty} \frac{B_{2i}}{(2i)!} \left[\frac{(p+2i-2)!}{(p-1)!(n-1)^{p+2i-1}} \right], \tag{C.1.1}$$

where $p = 4(1-H)$ which is between zero and one, and B_i is the i^{th} Bernoulli number. The Riemann zeta function $\zeta(p)$ is the infinite sum of $1/k^p$, and it can be shown that

$$\begin{aligned}
\zeta(p) &= \sum_{k=1}^{\infty} \frac{1}{k^p} \\
&= \frac{1}{1^p} + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \frac{1}{5^p} + \dots \\
&= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^p} + \frac{1}{2^{p-1}}\zeta(p) \\
(1 - 2^{1-p})\zeta(p) &= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^p} \\
\zeta(p) &= \frac{1}{1 - 2^{1-p}} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^p}.
\end{aligned} \tag{C.1.2}$$

This function is independent of n and its alternating series is convergent because the sequence decreases in absolute value which is asymptotic to zero as $k \to \infty$. Hence, this can be replaced by a constant. Then we expand the summation with Bernoulli number in Equation (C.1.1), i.e.,

$$\begin{aligned}
\sum_{\tau=1}^{n-1} \frac{1}{\tau^p} &\approx \zeta(p) - \frac{1}{(p-1)(n-1)^{p-1}} + \frac{1}{2(n-1)^p} \\
&\quad - \frac{B_2}{2!}\left[\frac{p!}{(p-1)!(n-1)^{p+1}}\right] - \frac{B_4}{4!}\left[\frac{(p+2)!}{(p-1)!(n-1)^{p+3}}\right] - \dots \\
\sum_{\tau=1}^{n-1} \frac{1}{\tau^p} &\approx \zeta(p) - \frac{1}{(p-1)(n-1)^{p-1}} + \frac{1}{2(n-1)^p} \\
&\quad - \frac{p}{12(n-1)^{p+1}} + \frac{p(p+1)(p+2)}{720(n-1)^{p+3}} - \dots
\end{aligned} \tag{C.1.3}$$

The first three terms with no Bernoulli number and the first two terms of the sequence with Bernoulli number are enough for the approximation of the p-series with good precision. This is because all other terms are very close to zero. In addition, since the alternating sequence with Bernoulli number decreases in absolute value as its order goes high, the p-series can be upper bounded by all terms shown in Equation (C.1.3)

such that

$$\sum_{\tau=1}^{n-1} \frac{1}{\tau^p} \leq \zeta(p) - \frac{1}{(p-1)(n-1)^{p-1}} + \frac{1}{2(n-1)^p} - \frac{p}{12(n-1)^{p+1}} + \frac{p(p+1)(p+2)}{720(n-1)^{p+3}}. \tag{C.1.4}$$

Substituting this upper boundary into Equation (7.2.2), it follows that

$$\frac{a_{max}^2}{n^2}\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] \leq \frac{4a_{max}^2}{n}\left\{s_0^2 + k^2\left[\zeta(p) - \frac{1}{(p-1)(n-1)^{p-1}} + \frac{1}{2(n-1)^p} - \frac{p}{12(n-1)^{p+1}} + \frac{p(p+1)(p+2)}{720(n-1)^{p+3}}\right]\right\}. \tag{C.1.5}$$

Because p is between zero and one, the term $4C/(n(p-1)(n-1)^{p-1})$ has the slowest decay comparing to the others. Thus, we can conclude that the variance of linear combinations of values of the periodogram is $\mathcal{O}(1/n^p)$ or $\mathcal{O}(1/n^{4(1-H)})$ when $0.75 \leq H < 1$.

Bibliography

R. K. Adenstedt. On large-sample estimation for the mean of a stationary random sequence. *Ann. Stat.*, 2(6):1095–1107, 1974.

R. J. Adler. Hausdorff dimension and Gaussian fields. *Ann. Probab.*, 5(1):145–151, 1977.

R. Akenhead, P. R. Hayes, K. G. Thompson, and D. French. Diminutions of acceleration and deceleration output during professional football match play. *J. Sci. Med. Sport*, 16(6):556–561, 2013.

S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollack. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3): 342–392, 2003.

A. A. Annis and E. H. Lloyd. The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika*, 63(1):111–116, 1976.

R. T. Baillie, T. Bollerslev, and H. O. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J. Econom.*, 74(1):3–30, 1996.

J. Bangsbo, F. M. Iaia, and P. Krustrup. Metabolic response and fatigue in soccer. *Int. J. Sports Physiol. Perform.*, 2(2):111–127, 2007.

J. M. Bardet, G. Lang, E. Moulines, and P. Soulier. Wavelet estimators of long-range dependent processes. *Stat. Inference Stoch. Process*, 3(1):85–99, 2000.

R. J. Barton and H. V. Poor. Signal detection in fractional Gaussian noise. *IEEE Trans. Inf. Theory*, 34(5):943–959, 1988.

J. Barunik and L. Kristoufek. On Hurst exponent estimation under heavy-tailed distributions. *Phys. A: Stat. Mech. Appl.*, 389(18):3844–3855, 2010.

J. Beran. *Statistics for long-memory processes*. Chapman & Hall/CRC Press, 1994.

J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. Commun.*, 43(2):1566–1579, 1995.

H. W. Borchers. *pracma: Practical numerical math functions*, 2022. URL `https://CRAN.R-project.org/package=pracma`. R package version 2.4.2.

P. S. Bradley, W. Sheldon, B. Wooster, P. Olsen, P. Boanas, and P. Krustrup. High-intensity running in English FA Premier League soccer matches. *J. Sports Sci.*, 27 (2):159–168, 2009.

P. J. Brockwell. *Time series: Theory and methods*. Springer-Verlag, 1987.

A. Contreras-Cristán, E. Gutiérrez-Peña, and S. G. Walker. A note on Whittle's likelihood. *Commun. Stat. - Simul. Comput.*, 35(4):857–875, 2006.

M. Corazza and A. G. Malliaris. Multi-fractality in foreign currency markets. *Multinatl. Financial J.*, 6(2):65–98, 2002.

G. Cornerllissen. Cosinor-based rhythmometry. *Theor. Biol. Med. Model.*, 11(16): 1–24, 2014.

M. Couillard and M. Davison. A comment on measuring the Hurst exponent of financial time series. *Phys. A: Stat. Mech. Appl.*, 348:404–418, 2005.

T. Di Matteo, T. Aste, and M. M. Dacorogna. Scaling behaviors in differently developed markets. *Phys. A: Stat. Mech. Appl.*, 324(1):183–188, 2003.

V. Di Salvo, W. Gregson, G. Atkinson, P. Tordoff, and B. Drust. Analysis of high intensity activity in Premier League soccer. *Int. J. Sports Med.*, 30(3):205–212, 2009.

V. Di Salvo, R. Baron, C. González-Haro, C. Gormasz, F. Pigozzi, and N. Bachl. Sprinting analysis of elite soccer players during European Champions League and UEFA Cup matches. *J. Sports Sci.*, 28(14):1489–1494, 2010.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.*, 74(366):427–431, 1979.

Z. Ding, C. W. J. Granger, and R. F. Eagle. A long memory property of stock market returns and a new model. *J. Empir. Finance*, 1(1):83–106, 1993.

B. Doucoure, K. Agbossou, and A. Cardenas. Time series prediction using artificial

wavelet neural network and multi-resolution analysis: Application to wind speed data. *Renew. Energy*, 92(1):202–211, 2016.

K. O. Dzhaparidze and S. Kotz. *Parameter estimation and hypothesis testing in spectral analysis of stationary time series*. Springer-Verlag, 1986.

R. H. T. Edwards. Biochemical bases of fatigue in exercise performance: Catastrophe theory of muscular fatigue. *In: Knuttgen HG, editor. Biochemistry of exercise. Champaign (IL): Human Kinetics*, 13(13):3–28, 1983.

T. Edwards, T. Spiteri, B. Piggott, J. Bonhotal, G. G. Haff, and C. Joyce. Monitoring and managing fatigue in basketball. *Sports*, 6(1):19, 2018.

C. Ellis and P. Wilson. Another look at the forecast performance of ARFIMA models. *Int. Rev. Financial Anal.*, 13(1):63–81, 2004.

K. Falconer. *Fractal geometry: Mathematical foundations and applications*. John Wiley & Sons Ltd., 1990.

M. Fernández-Martínez, J. L. García Guirao, M. A. Sánchez-Granero, and J. E. Trinidad Segovia. *Fractal Dimension for Fractal Structures: With Applications to Finance*. Springer Nature International Publishing, 2019.

R. Fossion, A. L. Rivera, J. C. Toledo-Roy, J. Ellis, and M. Angelova. Multiscale adaptive analysis of circadian rhythms and intradaily variability: Application to actigraphy time series in acute insomnia subjects. *PLOS ONE*, 12(7):e0181762, 2017.

R. Fox and M. S. Taqqu. Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Stat.*, 14(2):517–532, 1986.

K. Froggatt, S. Patel, G. Perez-Algorta, F. Bunn, G. Burnside, J. Coast, L. Dunleavy, C. Goodman, B. Hardwick, J. Kinley, N. Preston, and C. Walshe. Namaste Care in nursing care homes for people with advanced dementia: Protocol for a feasibility randomised controlled trial. *BMJ Open*, 8(11):e026531, 2018.

K. Froggatt, A. Best, F. Bunn, G. Burnside, J. Coast, L. Dunleavy, C. Goodman, B. Hardwick, C. Jackson, J. Kinley, A. Davidson-Lund, J. Lynch, P. Mitchell, G. Myring, S. Patel, G. Perez-Algorta, N. Preston, D. Scott, K. Silvera, and C. Walshe. A group intervention to improve quality of life for people with advanced dementia living in care homes: The Namaste feasibility cluster RCT. *Health Technol. Assess.*, 24(6):1–140, 2020.

J. Gatheral, T. Jaisson, and M. Rosenbaum. Volatility is rough. *Quant. Finance*, 18 (6):933–949, 2018.

T. Gneiting and M. Schlather. Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review*, 46(2):269–282, 2004.

B. S. B. Gonçalves, P. R. A. Cavalcanti, G. R. Tavares, T. F. Campos, and J. F. Araujo. Nonparametric methods in actigraphy: An update. *Sleep Sci.*, 7(3):158–164, 2014.

J. P. Grainger, A. M. Sykulski, P. Jonathan, and K. Ewans. Estimating the parameters of ocean wave spectra. *Ocean Eng.*, 229(1):108934, 2021.

C. W. J. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *J. Time. Ser. Anal.*, 1(1):15–29, 1980.

M. Griffin, G. Samorodnitsky, and D. S. Matteson. Likelihood inference for possibly non-stationary processes via adaptive overdifferencing. arXiv:2011.04168, 2011.

F. Halberg, Y. L. Tong, and E. A. Johnson. Circadian system phase – An aspect of temporal morphology; Procedures and illustrative examples. In *Von Mayersbach H. The cellular aspects of biorhythms*, pages 20–48. Heidelberg: Springer, 1967.

D. G. Harper, E. G. Stopa, A. C. McKee, A. Satlin, P. C. Harlan, R. Goldstein, and L. Volicer. Differential circadian rhythm disturbances in men with Alzheimer disease and frontotemporal degeneration. *Arch. Gen. Psychiatry*, 58(4):353–360, 2001.

C. F. Hatfield, J. Herbert, E. J. W. Van Someren, J. R. Hodges, and M. H. Hastings. Disrupted daily activity/rest cycles in relation to daily cortisol rhythms of home-dwelling patients with early Alzheimer's dementia. *Brain*, 127(5):1061–1074, 2004.

F. Hausdorff. Dimension und äußeres maß. *Math. Ann.*, 79(1):157–179, 1918.

A. M. Hooghiemstra, L. H. P. Eggermont, P. Scheltens, W. M. Van Der Flier, and E. J. A. Scherder. The rest-activity rhythm and physical activity in early-onset dementia. *Alzheimer Dis. Assoc. Disord.*, 29(1):45–49, 2015.

R. Hooke and T. A. Jeeves. "Direct Search" solution of numerical and statistical problems. *J. ACM.*, 8(2):212–229, 1961.

J. R. M. Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

K. Hu, D. G. Harper, S. A. Shea, E. G. Stopa, and F. A. J. L. Scheer. Noninvasive fractal biomarker of clock neurotransmitter disturbance in humans with dementia. *Sci. Rep.*, 3(2229):1–7, 2013.

K. Hu, R. F. R. Van Der Lek, M. Patxot, P. Li, S. A. Shea, F. A. J. L. Scheer, and E. J. W. Van Someren. Progression of dementia assessed by temporal correlations of physical activity: Results from a 3.5-year, longitudinal randomized controlled trial. *Sci. Rep.*, 6(27742):1–10, 2016.

S. E. Huber, P. Sachse, A. Mauracher, J. Marksteiner, W. Pohl, E. M. Weiss, and M. Canazei. Assessment of fractal characteristics of locomotor activity of geriatric in-patients with Alzheimer's dementia. *Front. Aging Neurosci.*, 11(272):1–16, 2019.

H. E. Hurst. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116:770–799, 1951.

E. A. F. Ihlen. Introduction to multifractal detrended fluctuation analysis in Matlab. *Front. Physiol.*, 3(141):1–18, 2012.

J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Phys. A: Stat. Mech. Appl.*, 316(1):87–114, 2002.

J. Karuppiah and C. A. Los. Wavelet multiresolution analysis of high-frequency Asian FX rates, Summer 1997. *Int. Rev. Financial Anal.*, 14(2):211–246, 2005.

S. S. Khan, B. Ye, B. Taati, and A. Mihailidis. Detecting agitation and aggression in people with dementia using sensors–A systematic review. *Alzheimers. Dement.*, 14 (6):824–832, 2018.

J. G. Kirkwood. Quantum statistics of almost classical assemblies. *Phys. Rev.*, 44(1): 31–37, 1933.

M. I. Knight and M. A. Nunes. Long memory estimation for complex-valued time series. *Stat. Comput.*, 29:517–536, 2019.

M. I. Knight, G. P. Nason, and M. A. Nunes. A wavelet lifting approach to long-memory estimation. *Stat. Comput.*, 27:1453–1471, 2017.

J. S. Kok, I. J. Berg, G. C. G. Blankevoort, and E. J. A. Scherder. Rest-activity rhythms in small scale homelike care and traditional care for residents with dementia. *BMC Geriatr.*, 17(137):1–8, 2017.

R. T. Krafty, H. Fu, J. L. Graves, S. A. Bruce, M. H. Hall, and S. F. Smagula. Measuring variability in rest-activity rhythms from actigraphy with application to characterizing symptoms of depression. *Stat. Biosci.*, 11:314–333, 2019.

P. Krustrp, M. Mohr, A. Steensberg, J. Bencke, M. Kjaer, and J. Bangsbo. Muscle and blood metabolites during a soccer game: Implications for sprint performance. *Med. Sci. Sports Exerc.*, 38(4):1165–1174, 2006.

N. A. Kyaw, C. A. Los, and S. Zong. Persistence characteristics of Latin American financial markets. *J. Multinatl. Financ. Manag.*, 16(3):269–290, 2006.

J. M. Lilly, A. M. Sykulski, J. J. Early, and S. C. Olhede. Fractional Brownian motion, the Matérn process, and stochastic modeling of turbulent dispersion. *Nonlin. Processes Geophys.*, 24(3):481–514, 2017.

A. W. Lo. Long-term memory in stock market prices. *Econometrica*, 59(5):1279–1313, 1991.

I. Lobato and P. M. Robinson. Averaged periodogram estimation of long memory. *J. Econom.*, 73(1):303–324, 1996.

M. Lyons, Y. Al-Nakeeb, and A. Nevill. The impact of moderate and high intensity total body fatigue on passing accuracy in expert and novice basketball players. *J. Sports Sci. Med.*, 5(2):215–227, 2006.

S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989.

K. S. Man. Long memory time series and short term forecasts. *Int. J. Forecast.*, 19 (3):477–491, 2003.

B. B. Mandelbrot. Self-affine fractals and fractal dimension. *Phys. Scr.*, 32(4):257–260, 1985.

B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

A. I. McLeod and K. W. Hipel. Preservation of the rescaled adjusted range: 1. A reassessment of the Hurst phenomenon. *Water Resour. Res.*, 14(3):491–508, 1978.

M. Mohr, P. Krustrup, and J. Bangsbo. Match performance of high-standard soccer players with special reference to development of fatigue. *J. Sports Sci.*, 21(7):519–528, 2003.

M. Mohr, P. Krustrup, and J. Bangsbo. Fatigue in soccer: A brief review. *J. Sports Sci.*, 23(6):593–599, 2005.

F. J. Molz and H. H. Liu. Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. *Water Resour. Res.*, 33(10):2273–2286, 1997.

M. J. Mullaney, M. P. McHugh, T. M. Donofrio, and S. J. Nicholas. Upper and lower extremity muscle fatigue after a baseball pitching performance. *Am. J. Sports Med.*, 33(1):108, 113 2005.

E. S. Musiek, M. Bhimasani, M. A. Zangrilli, J. C. Morris, D. M. Holtzman, and Y. S. Ju. Circadian rest-activity pattern changes in aging and preclinical Alzheimer disease. *JAMA Neurol.*, 75(5):582–590, 2018.

T. A. Øigård, A. Hanssen, and L. L. Scharf. Spectral correlations of fractional Brownian motion. *Phys. Rev. E*, 74(3):1–6, 2006.

J. M. Oliva-Lozano, V. Fortes, P. Krustrup, and J. M. Muyor. Acceleration and sprint profiles of professional male football players in relation to playing position. *PLOS ONE*, 15(8):e0236959, 2020.

J. M. Oosterman, B. Van Harten, R. L. Vogels, A. A. Gouw, H. C. Weinstein, P. Schel-

tens, and E. J. A. Scherder. Distortions in rest–activity rhythm in aging relate to white matter hyperintensities. *Neurobiol. Aging*, 29(8):1265–1271, 2008.

V. Paxson. Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. *Comput. Commun. Rev.*, 27(5):5–18, 1997.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49(2):1685–1689, 1994.

D. B. Percival and A. T. Walden. *Spectral analysis for physical applications*. Cambridge: Cambridge University Press, 1993.

D. B. Percival and A. T. Walden. *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press, 2000.

B. Qian and K. Rasheed. Hurst exponent and financial market predictability, 2007.

E. Rampinini, F. M. Impellizzeri, C. Castagna, A. J. Coutts, and U. Wisløff. Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue and competitive level. *J. Sci. Med. Sport*, 12(1):227–233, 2009.

E. Rampinini, A. Bosio, I. Ferraresi, A. Petruolo, A. Morelli, and A. Sassi. Match-related fatigue in soccer players. *Med. Sci. Sports Exerc.*, 43(11):2161–2170, 2011.

T. Reilly. Energetics of high-intensity exercise (soccer) with particular reference to fatigue. *J. Sports Sci.*, 15(3):257–263, 1997.

A. W. Rihaczek. Signal energy distribution in time and frequency. *IEEE Trans. Inf. Theory*, 14(3):369–374, 1968.

P. M. Robinson. *Time series with long memory.* Oxford: Oxford University Press, 2003.

B. G. Sanderson, A. Goulding, and A. Okubo. The fractal dimension of relative lagrangian motion. *Tellus A: Dyn. Meteorol. Oceanogr.*, 42(5):550–556, 1990.

A. Satlin, L. Volicer, E. G. Stopa, and D. G. Harper. Circadian locomotor activity and core-body temperature rhythms in Alzheimer's disease. *Neurobiol. Aging*, 16 (5):765–771, 1995.

A. Sensoy. Generalized Hurst exponent approach to efficiency in MENA markets. *Phys. A: Stat. Mech. Appl.*, 392(20):5019–5026, 2013.

J. R. Silva, M. C. Rumpf, M. Hertzog, C. Castagna, A. Farooq, O. Girard, and K. Hader. Acute and residual soccer match-related fatigue: A systematic review and meta-analysis. *Sports Med.*, 48(3):539–538, 2018.

S. F. Smagula, S. Gujral, C. S. Capps, and R. T. Krafty. A systematic review of evidence for a role of rest-activity rhythms in dementia. *Front. Psychiatry*, 10(778): 1–7, 2019.

K. Suibkitwanchai, A. M. Sykulski, G. Perez-Algorta, D. Waller, and C. Walshe. Non-parametric time series summary statistics for high-frequency accelerometry data from individuals with advanced dementia. *PLOS ONE*, 15(9):e0239368, 2020.

A. M. Sykulski, S. C. Olhede, A. P. Guillaumin, J. M. Lilly, and J. J. Early. The debiased Whittle likelihood. *Biometrika*, 106(2):251–266, 2019.

M. S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: An empirical study. *Fractals*, 3(4):785–798, 1995.

C. Twist and J. Highton. Monitoring fatigue and recovery in rugby league players. *Int. J. Sports Physiol. Perform.*, 8(5):467–474, 2013.

V. T. Van Hees, L. Gorzelniak, E. C. Dean León, M. Eder, M. Pias, S. Taherian, U. Ekelund, F. Renström, P. W. Franks, A. Horsch, and S. Brage. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLOS ONE*, 8(4):e0061691, 2013.

E. J. W. Van Someren, J. M. Oosterman, B. Van Harten, R. L. Vogels, A. A. Gouw, H. C. Weinstein, A. Poggesi, P. Scheltens, and E. J. A. Scherder. Medial temporal lobe atrophy relates more strongly to sleep-wake rhythm fragmentation than to age or any other known risk. *Neurobiol. Learn. Mem.*, 160:132–138, 2019.

M. Waldron, J. Highton, M. Daniels, and C. Twist. Preliminary evidence of transient fatigue and pacing during interchanges in rugby league. *Int. J. Sports Physiol. Perform.*, 8(2):157–164, 2013.

Y. Z. Wang, B. Li, R. Q. Wang, J. Su, and X. X. Rong. Application of the Hurst exponent in ecology. *Comput. Math. with Appl.*, 61(8):2129–2131, 2011.

P. Whittle. Estimation and information in stationary time series. *Ark. Mat.*, 2(5): 423–434, 1953.

W. Witting, I. H. Kwa, P. Eikelenboom, M. Mirmiran, and D. F. Swaab. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. *Biol. Psychiatry*, 27(6):563–572, 1990.

J. Xiu and Y. Jin. Empirical study of ARFIMA model based on fractional differencing. *Phys. A: Stat. Mech. Appl.*, 377(1):138–154, 2007.

Y. Yajima. On estimation of long-memory time series models. *Aust. N. Z. J. Stat.*, 27(3):303–320, 1985.

L. Zhang, P. A. Mykland, and Y. Aït-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *J. Am. Stat. Assoc.*, 100(472): 1394–1411, 2005.

L. A. Zuurbier, M. A. Ikram, A. I. Luik, A. Hofman, E. J. W. Van Someren, M. W. Vernooij, and H. Tiemeier. Cerebral small vessel disease is related to disturbed 24-h activity rhythms: A population-based study. *Eur. J. Neurol.*, 22(11):1482–1487, 2015.