

A Study of Chaotic Intermittency Maps and An Analysis of Consumer Data

D.J. Natsios

November 4, 2009

A Study of Chaotic Intermittency Maps and An Analysis of Consumer Data

D.J. Natsios

Abstract: This Thesis consists of two distinct parts. The first is concerned with the study of the long memory characteristics of a relatively new class of chaotic intermittency maps. In particular we study the symmetric cusp map, the asymmetric cusp map, polynomial maps and logarithmic maps. In previous studies by Bhansali and Holland, it has been shown that these maps can simulate stationary time series with a full range of values for the long memory parameter, including $d = 0.5$ which is usually considered non-stationary, $d = 0$ which is usually considered short memory and $d < 0$ which is usually intermediate memory. Further more, for each given map with a given set of parameters, asymptotic proofs are available to give the 'true' value of d to be taken as known. This gives us the opportunity to carry out a simulation study to test various long memory estimation techniques, namely the GPH method, the Local Whittle Method, the FExp method and the FAR method (using both the BIC and AIC), when the standard assumptions of linearity and Gaussian distribution no longer hold.

Our results, based on 1000 simulations of series length 10,000 for each of 12 maps, help to reconfirm the asymptotic expectations of d , although we show that bias increases considerably near the boundary conditions of 0 and 0.5 and that to remove these biases we may need to increase our series length to over a million observations.

This gives motivation into the study of a new dual parameter model for long memory. Theoretical work is given to show the biases of standard estimation techniques can be significant when an unbounded slowly varying function is present, and new extensions to the GPH and Local Whittle estimation techniques to include this term are provided. A fully parameterised extension to the FARIMA model is also presented and simulation studies are carried out to justify the use of this model alongside the FARIMA model.

In the second part, we carry out an analysis on some new real human motion data provided by Unilever. This data concerns the movement of seven sensors attached to a human subject whilst applying a deodorant stick to their underarm, the raw data being the (x,y,z) co-ordinates with time stamp of the sensors. We look into methods of reducing dimensions with the application of principal component analysis. We then provide easier to interpret transformations and group the data in different application techniques. Finally, the data is modelled with the use of B-Splines and piecewise Bezier curves with the ultimate aim of simulating further life like results.

CONTENTS

<i>Part I</i>	5
1. <i>Introduction</i>	6
1.1 Stationarity	8
1.2 Hilbert Spaces	9
1.3 Wold Decomposition Theorem	10
1.4 Best Linear Predictor	11
1.5 Autocorrelation Function and Partial Autocorrelation Function	12
1.6 ARMA Models	15
1.6.1 Linear Filters	19
1.6.2 Sample Mean and ACF	21
1.6.3 Parameter Estimation	23
1.6.4 Model Selection	31
1.7 The Frequency Domain	34
1.7.1 Fourier Analysis	35
1.7.2 Spectral Density	37
1.7.3 Periodogram	39
1.7.4 SGF, Gain and Phase	42
1.8 Exp Model	43
2. <i>Long-Memory Processes</i>	46
2.1 Definition of a Long-Memory Process	46
2.2 Long-Memory Models	48
2.2.1 Fractional Gaussian Noise	48
2.2.2 FARIMA model	50
2.2.3 FExp model	53
2.3 Estimation of the Long Memory Parameter	55
2.3.1 FARIMA Method	55
2.3.2 Geweke and Porter-Hudak (GPII) method	58
2.3.3 The Local Whittle Method	61
2.3.4 The FExp method	63
2.3.5 FAR Method	65
3. <i>Chaos</i>	68
3.1 Basic Properties	69
3.1.1 Sensitivity Dependence	69

3.1.2	Transitivity	71
3.1.3	Periodic Points	72
3.1.4	Invariant Density	75
3.2	Chaotic Intermittency Maps	77
3.2.1	The Cusp Maps	78
3.2.2	The Polynomial Maps	81
3.2.3	The Logarithmic Maps	82
4.	<i>Estimation of the Correlation Decay Rates for Chaotic Intermittency Maps</i>	85
4.1	Simulation Study	85
4.1.1	Plan of Study	89
4.1.2	Simulation Results	90
4.2	Further Simulation Studies	100
4.2.1	Extended Length Simulation Study	102
4.2.2	Systematic Sampling Simulation Study	106
4.3	Conclusions	112
5.	<i>Dual Parameter Long Memory Model</i>	116
5.1	Dual Parameter GPII Method	119
5.1.1	Technical Lemmas	135
5.2	Dual Parameter Local Whittle Method	137
5.2.1	Technical Lemmas	148
5.3	Dual Parameter FARIMA model	169
5.4	Simulation Study	177
5.4.1	Empirical Results of the Non-Parametric Methods	178
5.4.2	Empirical Results for Fitting DFARIMA Models	191
5.4.3	Application to Other Series	200
5.4.4	Application to Chaotic Intermittency Maps	209
6.	<i>Stochastic Intermittency Maps</i>	218
6.1	Stochastic Polynomial Map	219
6.1.1	Estimation of Alpha for the Stochastic Polynomial Map	221
6.2	Stochastic Logarithmic Map	223
6.2.1	Estimation of Beta for the Stochastic Logarithmic Map	225
6.3	Stochastic Polynomial-Logarithmic Map	226
6.3.1	Simultaneous Estimation of Alpha and Beta	228
6.4	Empirical Studies	229
6.4.1	Laminar Region Behaviour	230
6.4.2	Estimation of Alpha and Beta	236
6.4.3	Discrimination between the Maps	243

<i>Part II</i>	251
7. <i>Deodorant Stick Data Analysis</i>	252
7.1 Data Collection	253
7.2 Under Arm Data	255
7.2.1 Change Point Analysis	255
7.2.2 ANOVA Application	260
7.2.3 Frequency Domain ANOVA Analysis	263
7.3 Multivariate Analysis	270
7.3.1 Principal Component Analysis	270
7.3.2 PCA of Y_{ijk}	273
7.3.3 Graphical Representation	280
7.4 Modelling and Simulation	284
7.4.1 Grouping	284
7.4.2 Vector AR Models	287
7.4.3 B-Splines and Bezier Curves	298
7.5 Summary	313

Part I

1. INTRODUCTION

This Thesis consists of two distinct parts. The first part is concerned with the study of the long memory characteristics of a relatively new class of chaotic intermittency maps, and the new extensions to long memory estimation techniques and models that the study of these chaotic maps motivates.

The remainder of Chapter 1 introduces basic key concepts of time series analysis such as stationarity, the Wold decomposition theorem, the autocorrelation function and the frequency domain. It also introduces the commonly used ARMA model and the Exp models of Bloomfield [1973].

The fundamental ideas of Chapter 1 are then extended in Chapter 2 to introduce long memory processes. Long memory time series analysis is becoming increasingly popular in time series literature. Examples of application can be found from varying fields such as internet traffic and environmental issues. Early examples can be seen in Hurst [1951],[1957], Mandelbrot and Wallis [1968], Mandelbrot [1972] and McLeod and Hipel [1978] amongst others. Despite this increasing popularity, however, there is no universally agreed definition of long memory and Chapter 2 gives some of the more commonly used definitions, settling on the two key aspects of stationary and an ACF that is not absolutely summable to define long memory for this thesis.

The ARMA and Exp models of Chapter 1 are extended to the long memory FARIMA and FExp models, defined by Granger and Joyeux [1980] and Beran [1993] respectively. Details of several methods that have been developed for estimating the long memory parameter associated with these models are also given, along with semi- and non-parametric methods.

These estimators are usually tested for bias, consistency and robustness in simulation studies involving linear long memory models, see, for example Taqqu and Teverovsky [1996]. Following the work of Bhansali and Holland [2008b], chaotic intermittency maps are used here as an alternative to these linear models. The concept of chaos is introduced in Chapter 3, along with descriptions of some fundamental properties. Evidence of chaos has been found in a wide variety of sources, such as acoustics, chemistry, circuits, lasers and plasmas, see Lauterborn [1981], Simoyi et al [1982], Rollins and Hunt [1984], Mork et al [1990] and Sagdeev et al [1990], respectively. The class of chaotic intermittency maps of interest, defined in section 3.2, have previously been considered in the modelling of internet traffic, see Bhansali et al [2005].

The new work begins in Chapter 4 with an investigative simulation study on the

rate of decay of the correlations of the series produced by these chaotic intermittency maps. The chaotic intermittency maps studied here are the symmetric cusp map, the asymmetric cusp map, polynomial maps and logarithmic maps. In previous studies by Bhansali and Holland [2008b], it has been shown that these maps can simulate stationary time series with a full range of values for the long memory parameter, including $d = 0.5$ which is usually considered non-stationary, $d = 0$ which is usually considered short memory and $d < 0$ which is usually intermediate memory. Furthermore, for each given map with a given set of parameters, asymptotic proofs are available to give the 'true' value of d to be taken as known. This gives the opportunity to carry out a simulation study to test various long memory estimation techniques, namely the GPII method, the Local Whittle Method, the FExp method and the FAR method, introduced in Chapter 2, when the assumptions of linearity and Gaussian distribution no longer hold.

The results of Chapter 4, which only partly agree with the asymptotic theory, motivates the new dual parameter long memory model introduced in Chapter 5. Section 5.3 introduces a new extension to the FARIMA model which allows for a second long memory parameter. This model allows for both 'weak' and 'strong' long memory boundary behaviour and includes the FARIMA model as a special case.

Theoretical work is carried out to give the asymptotic distributions of all the newly defined estimators. The proofs for the newly defined extended versions of the GPII and Local Whittle methods follow those of Robinson [1995a] and Robinson [1995b] and are presented in a level of detail similar to those papers. An application of these new methods to the chaotic intermittency maps is carried out and shown to improve the results. In addition, simulation studies show these new methods can also outperform older methods in more standard linear Gaussian cases, and an application to Bellcore Ethernet data shows the fit of the new DFARIMA model may be preferable to the standard FARIMA model.

Finally for Part 1, new stochastic extensions of these chaotic intermittency maps are introduced in Chapter 6. The study of stochastic versions of chaotic maps has been carried out previously by several authors, for example, Chan and Tong [1994], [2001], Alves and Arujo [2000], Alves and Viana [2002] and Alves et al [2004].

The new stochastic versions of the intermittency maps introduced here are shown to possess the same properties as the deterministic maps in the laminar region and hence retain the asymptotic rates of decay discussed in section 3.2. Simulation studies are then carried out to reconfirm these findings and use the results to estimate the parameters of the maps. The forms of the stochastic maps presented here were suggested by Dr M. Holland and have been included and studied in this chapter with his permission.

The second part of this thesis focuses on the analysis of consumer data provided by Unilever. The data is concerned with the movement of individuals over time and space whilst applying a deodorant stick to the area under their left arm, the raw data being the (x, y, z) co-ordinates with time stamp of seven sensors attached to

the individual.

The deodorant stick data being discussed in this paper has been collected and supplied by Unilever and is part of a larger series of experiments being carried out using the company's recently acquired motion sensor technology. Motion sensor technology has been used previously in several areas, such as virtual simulations of factories and equipment to test and improve designs, see for example Faraway and Reed [2007], as well as fields such as sport and medicine, see Menache [2000].

A description of the deodorant stick data, including how it was collected and recorded is given in section 7.1. It would be of interest to Unilever to discover if such an experiment can pick up differences between individuals and products and group individuals and products into different application techniques and categories. The cost of various experiments at Unilever could also be greatly reduced if suitable methods of modelling the data and simulating new data could be found.

Section 7.3 attempts to better interpret the data by application of principal component analysis. A new form of the data is then derived in section 7.3.3. This new transformed form of the data provides a reduction in dimensionality without loss of information whilst simultaneously providing easier to understand definitions to the series.

Finally, section 7.4 looks at methods of modelling the data in this new form and simulating new data with the same properties as the original. Section 7.4.2 introduces the use of vector autoregressive models, whilst section 7.4.3 looks instead at modelling the data by fitting functions such as B-Splines and Bezier curves.

1.1 Stationarity

A stochastic process $\{y_t\} \forall t \in \mathbb{Z}$ is said to be *weakly* or *second-order stationary* if it meets the following three conditions:

1. The $E(y_t)$ must exist and be independent of $t \forall t \in \mathbb{Z}$,

$$E(y_1) = E(y_2) = E(y_3) = \dots = \mu. \quad (1.1)$$

2. The $Var(y_t)$ must also exist and be independent of $t \forall t \in \mathbb{Z}$,

$$Var(y_t) = E((y_t - \mu)^2) = \sigma^2 \quad (1.2)$$

3. The $Cov(y_t, y_{t-u})$ must be independent of $t, \forall t, u \in \mathbb{Z}$, and therefore a function of u .

$$Cov(y_t, y_{t-u}) = E((y_t - \mu)(y_{t-u} - \mu)) = R(u) \quad (1.3)$$

A process $\{y_t\} \forall t \in \mathbb{Z}$ is said to be *strictly stationary* if and only if the joint distribution of y_1, y_2, \dots, y_k is identical to the joint distribution of $y_{1+h}, y_{2+h}, \dots, y_{k+h}$ for all integers h and k . A time series can be weakly stationary without being strictly stationary and can be strictly stationary without being weakly stationary.

Hereafter, a stationary process will always refer to a weakly stationary or second-order stationary process unless otherwise specified.

1.2 Hilbert Spaces

Let V be a vector space. A *norm* on the vector space is a function $\|\cdot\| : V \rightarrow [0, \infty)$ such that for all $x, y \in V$,

$$\|x\| = 0 \Leftrightarrow x = 0,$$

$$\|Cx\| = |C| \|x\|, \quad \text{for all } C \in \mathbb{R}$$

and

$$\|x + y\| \leq \|x\| + \|y\|.$$

The sequence $\{x_t\} \in V$ is called a *Cauchy sequence* if and only if for all $\varepsilon > 0$ there exists an integer n such that

$$\|x_t - x_s\| < \varepsilon \quad \text{for all } t, s > n.$$

If every Cauchy sequence has a limit in the vector space, V , then V is said to be *complete*.

An *inner product* is a function $\langle \cdot, \cdot \rangle : V \rightarrow \mathbb{R}$ such that for all $x, y, z \in V$,

$$\langle x, x \rangle \geq 0, \tag{1.4}$$

$$\langle x, x \rangle = 0 \Leftrightarrow x = 0, \tag{1.5}$$

$$\langle C_1x + C_2y, z \rangle = C_1\langle x, z \rangle + C_2\langle y, z \rangle, \quad \text{for all } C_1, C_2 \in \mathbb{R} \tag{1.6}$$

and

$$\langle x, y \rangle = \langle y, x \rangle. \tag{1.7}$$

The vector space, V , is an *inner product space* if it is endowed with an inner product $\langle \cdot, \cdot \rangle$. For an inner product, $\langle \cdot, \cdot \rangle$, there exists a norm $\|x\| = \sqrt{\langle x, x \rangle}$ in V .

An example of an inner product is the autocovariance for a stationary process $\{y_t\}$ with $\mu = 0$. Since $\text{Cov}(y_t, y_t) = \text{Var}(y_t)$ and $\text{Cov}(y_t, y_{t-u}) = \text{Cov}(y_{t-u}, y_t)$, for all $t, u \in \mathbb{Z}$, equations 1.4, 1.5 and 1.7 are satisfied. Also, for all $t, u, k \in \mathbb{Z}$,

$$\begin{aligned} \text{Cov}(C_1y_t + C_2y_{t-u}, y_{t-k}) &= E\{(C_1y_t + C_2y_{t-u})y_{t-k}\} - E(C_1y_t + C_2y_{t-u})E(y_{t-k}) \\ &= E(C_1y_t y_{t-k}) - E(C_1y_t)E(y_{t-k}) + E(C_2y_{t-u} y_{t-k}) - E(C_2y_{t-u})E(y_{t-k}) \end{aligned}$$

$$= C_1 \text{Cov}(y_t, y_{t-k}) + C_2 \text{Cov}(y_{t-u}, y_{t-k}),$$

which satisfies equation 1.6. Note, this gives the standard deviation as a norm, since $\sigma_y = \sqrt{\text{Var}(y_t)}$.

A *Hilbert space*, H , is a complete inner product vector space. The following *projection theorem* concerning Hilbert spaces can be found in Palma [2007], amongst others.

Theorem 1.2.1. *Let H be a Hilbert space, let $V \subset H$ be a closed subspace and let $x \in H$. Then,*

a) *there is unique point $y \in H$ such that*

$$\|x - y\| = \inf_{z \in V} \|x - z\|$$

b) *for $y \in V$,*

$$\|x - y\| = \inf_{z \in V} \|x - z\| \Leftrightarrow \langle x - y, z \rangle = 0, \quad \text{for all } z \in V.$$

Given a closed subset, V , of a Hilbert space, H , the space generated by all finite linear combinations of elements of V is called the *span* of V and is denoted by $sp(V)$. The space containing all the limits of the sequences in $sp(V)$ is denoted by $\overline{sp}(V)$ and is the closure of $sp(V)$ in H .

1.3 Wold Decomposition Theorem

Let $F_t^s = \overline{sp}(\{y_s\})$ with $s < t$, that is, F_t^s contains all the past information of the process $\{y_t\}$ at time t . The process is said to be *linearly deterministic*, or *singular*, if and only if $F_{-\infty}^t = \dots = F_{-\infty}^s = F$ independent of t . This implies that all available information was available at a point infinitely into the past and all values since then have been generated from this. Thus the process is perfectly predictable. Alternatively, a process is said to be *purely nondeterministic* or *regular* if and only if $F_{-\infty}^t = \{0\}$.

The following theorem is known as the *Wold Decomposition Theorem*, the proof of which can be seen in Wold [1953].

Theorem 1.3.1. *Any stationary process is the sum of two orthogonal processes such that one is singular and the other is regular. Furthermore, this decomposition is unique and the regular part may be expressed as*

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \tag{1.8}$$

where $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2 \forall t \in \mathbb{Z}$, and $E(\varepsilon_t \varepsilon_s) = 0$ for $s \neq t$, $t, s \in \mathbb{Z}$.

Take $\{y_t\}$ and $\{\varepsilon_t\}$ to be defined as in Theorem 1.3.1. The process $\{y_t\}$ is said to be *causal* since it is dependent only on the present and the past values of $\{\varepsilon_t\}$. Causality is assumed throughout this thesis. Now, since $E(\varepsilon_t) = 0 \forall t \in \mathbb{Z}$, this gives $E(y_t) = 0 \forall t \in \mathbb{Z}$ and the covariances of y_t at lag u are given by

$$\text{Cov}(y_t, y_{t-u}) = E(y_t y_{t-u}) = E \left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-u-j} \right).$$

Since the $\{\varepsilon_t\}$ are uncorrelated, this gives

$$\text{Cov}(y_t, y_{t-u}) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+u} = R(u), \quad (1.9)$$

which are finite and independent of $t \forall u \in \mathbb{Z}$, due to the condition $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

The sequence $\{\varepsilon_t\}$ is known as a *white noise* process. If ε_t and ε_s , $s \neq t$, are independent and identically distributed random variables $\forall s, t \in \mathbb{Z}$ then $\{\varepsilon_t\}$ is a *strict white noise* process. Note that the Wold Decomposition Theorem requires only that the sequence $\{\varepsilon_t\}$ be *uncorrelated* not independent. Such a series is known as a *weak white noise* process.

1.4 Best Linear Predictor

Let $\{y_t\}$ and $\{\varepsilon_t\}$ to be defined as in Theorem 1.3.1. The process $\{y_t\}$ is said to be *invertible* if there exists a sequence of coefficients $\{\pi_j\}$ such that

$$\varepsilon_t = - \sum_{j=0}^{\infty} \pi_j y_{t-j},$$

where $\pi_0 = -1$, or equivalently,

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} \pi_j y_{t-j}.$$

Assuming $\{y_t\}$ is invertible, the best linear prediction of y_t based on its history $\bar{s}p(y_{t-1}, y_{t-2}, \dots)$ is given by

$$\hat{y}_t = \sum_{j=1}^{\infty} \pi_j y_{t-j},$$

and $\varepsilon_t = y_t - \hat{y}_t$, is an orthogonal process representing the part of y_t that cannot be linearly predicted from the past.

Now, given a finite observed series $\{y_t : t \in (1, \dots, n)\}$, the best linear predictor of y_{n+1} based on its finite past $\bar{s}p(y_1, \dots, y_n)$ is given by

$$\hat{y}_{n+1} = \sum_{j=0}^n \phi_j y_{n-j},$$

for some $\phi_0, \dots, \phi_n \in \mathbb{R}$. Theorem 1.2.1 gives

$$\langle y_{n+1} - \hat{y}_{n+1}, y_j \rangle = 0, \quad \text{for } j \in (1, \dots, n),$$

and from the property of inner products given in equation 1.6, the coefficients $\{\phi_j\}$ satisfy

$$\sum_{i=1}^n \phi_i \langle y_{n+1-i}, y_j \rangle = \langle y_{n+1}, y_j \rangle, \quad \text{for } j \in (1, \dots, n).$$

If $\{y_t\}$ is a stationary mean zero process, taking the inner product to be the autocovariances gives,

$$\sum_{i=1}^n \phi_i R(n+1-i-j) = R(n+1-j), \quad \text{for } j \in (1, \dots, n).$$

Estimation of these coefficients and autocovariances from a given time series is discussed in sections 1.6.3 and 1.6.2.

1.5 Autocorrelation Function and Partial Autocorrelation Function

For a stationary process $\{y_t\}$, the function $R(u)$, see equation 1.3, is known as the *autocovariance function*, with $R(0)$ equal to the variance. Correlation between y_t and y_{t-u} is defined as

$$\text{cor}(y_t, y_{t-u}) = \frac{\text{Cov}(y_t, y_{t-u})}{\sqrt{\text{var}(y_t)\text{var}(y_{t-u})}}.$$

Since $\{y_t\}$ is stationary, $\text{var}(y_t) = \text{var}(y_{t-u}) = R(0)$ and $\text{Cov}(y_t, y_{t-u}) = R(u)$, this gives rise to the *autocorrelation function* (ACF)

$$\text{cor}(y_t, y_{t-u}) = \frac{R(u)}{R(0)} = r(u). \quad (1.10)$$

The autocorrelation function is often used instead of the autocovariance function since, by definition, $r(0) = 1$ for any process. Some other basic properties of the

autocorrelation and autocovariance functions are given in the following Theorem, see Box and Jenkins [1976] for a proof.

Theorem 1.5.1. For a stationary process $\{y_t\}$, $t \in \mathbb{Z}$, with variance $R(0) \neq 0$, the autocovariance function, $R(u)$, given by equation 1.3, and the autocorrelation function, $r(u)$, given by equation 1.10, have the following properties:

a) The autocorrelation (or autocovariance) function is an even function:

$$r(u) = r(-u)$$

b) The autocorrelation (or autocovariance) function is positive-definite for any stationary process:

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j r(|i-j|) > 0 \quad c_i, c_j \in \mathbb{R}, n \in \mathbb{N},$$

such that at least one $c_i \neq 0$, i in $(1 \dots n)$.

c) The autocorrelation (or autocovariance) function at lag u is such that

$$|r(u)| \leq r(0).$$

The plot of the autocorrelations against its lag u is known as the *correlogram*. Inspection of the correlogram may give a general idea of the type of model that should be fitted to the data. This is discussed further in section 1.6. Larger values of the ACF at higher lags implies the present value of y_t is more strongly linearly dependent on distant past values. The rate at which the correlogram tends to zero thus gives an indication of the *memory* of the process, which is central to the idea of Long Memory, discussed in Chapter 2. Plotting covariances would give the same shape as the correlogram, but plotting correlations gives a standardised measure.

The *partial autocorrelation function* (PACF), $\Phi(u)$, is the autocorrelation between y_t and y_{t-u} that is not accounted for by lags 1 to $u-1$. That is, it is a measure of the linear connection between y_t and y_{t-u} when the linear influence of the random variables that lie between have been filtered out, see for example, Brockwell and Davis [1991]. It is defined, for $u \geq 2$, by

$$\Phi(u) = \text{Corr}(y_t - \mathcal{P}(y_t|y_{t-1}, \dots, y_{t-u+1}), y_{t-u} - \mathcal{P}(y_{t-u}|y_{t-1}, \dots, y_{t-u+1})) \quad (1.11)$$

where $\mathcal{P}(y|X)$ is the best linear projection of y on X , see section 1.4. For $u=1$, $\Phi(u) = r(u)$, since the projection onto $\{0\}$ is $\{0\}$. The following theorem, due to Ramsay [1974], gives the properties of a PACF for a stationary process.

Theorem 1.5.2. For a stationary process $\{y_t\}$, $t \in \mathbb{Z}$, the PACF, $\Phi(u)$, given by equation 1.11, has the following properties:

a) The PACF is an even function:

$$\Phi(u) = \Phi(-u)$$

b) The PACF is non negative-definite for any stationary process:

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j \Phi(|i-j|) \geq 0 \quad \forall c_i, c_j \in \mathbb{R}, n \in \mathbb{N}$$

c) The PACF at lag u is such that

$$|\Phi(u)| \leq \Phi(0) = 1.$$

The asymptotic behaviour of the PACF has been studied by Inoue [2000], in which he gives the following theorem.

Theorem 1.5.3. *Let $\{y_t\}$ be a stationary process with Wold Decomposition defined by equation 1.8, with $\psi_j \geq 0 \forall j \in \mathbb{Z}$, then*

$$|\Phi(u)| \sim \frac{r(u)}{\sum_{j=-u}^u r(j)}$$

as $u \rightarrow \infty$.

Given the ACF, the Durbin-Levinson algorithm can be used to find the PACF, see Durbin [1960] and Levinson [1947]. The algorithm starts by defining $\phi_{11} = r(1)$ and $\nu_1 = (1 - r^2(1))$, then using the recursive equations

$$\phi_{mm} = \frac{r(m) - \sum_{j=1}^{m-1} \phi_{m-1,j} r(m-j)}{\nu_{m-1}}, \quad (1.12)$$

$$\begin{pmatrix} \phi_{m1} \\ \phi_{m2} \\ \vdots \\ \phi_{m,m-1} \end{pmatrix} = \Phi_{m-1} - \phi_{mm} \begin{pmatrix} \phi_{m-1,m-1} \\ \phi_{m-1,m-2} \\ \vdots \\ \phi_{m-1,1} \end{pmatrix} \quad (1.13)$$

and

$$\nu_m = \nu_{m-1}(1 - \phi_{mm}^2), \quad (1.14)$$

where $\Phi_{m-1} = (\phi_{m-1,1}, \phi_{m-1,2}, \dots, \phi_{m-1,m-1})'$. Note, from equation 1.12, ϕ_{mm} is the correlation between y_t and y_{t-m} after the correlation on $y_{t-1}, \dots, y_{t-m+1}$ has been removed. Hence, the PACF is given by $\Phi(u) = \phi_{uu}$.

Note that given the PACF, these recursive equations can be solved to find the ACF. Hence the PACF contains all the information of the ACF, and vice versa, and it is possible to characterise a stationary process by either its PACF or its ACF.

1.6 ARMA Models

The use of *Autoregressive Moving Average* (ARMA) models are very popular in the literature. Their properties were extensively studied by Box and Jenkins [1976] and computational packages for simulation, estimation and prediction of ARMA processes are now available for a large range of software, such as S and R. As such, they have been applied to a wide variety of actual data, for example the much studied sunspot data, wheat prices, magnetic resonance image reconstruction etc see Whittle [1954], Sargan [1953] and Smith et al [1986]. Discussion of these models are widely available in the literature and reviews can be seen, for example, by Anderson [1971], Brockwell and Davis [1991], Hannan [1980] and Wei [1989].

Starting from the Wold representation, see Theorem 1.3.1, a stationary process $\{y_t\}$ with mean μ and no other deterministic part can be written as

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

where $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2 \forall t \in \mathbb{Z}$, and $E(\varepsilon_t \varepsilon_s) = 0$ for $s \neq t$. Note, μ here represents the singular part of the process.

A *Moving Average*, MA(q), process is one in which this infinite summation is truncated at lag q, that is $\psi_j = 0 \forall j > q$, thus

$$y_t = \mu + \sum_{j=0}^q \psi_j \varepsilon_{t-j}.$$

For ease of notation, introduce now the *lag function*, L. The Lag function, L, is defined by the property

$$Ly_t = y_{t-1}$$

The function is assumed to have $|L| < 1$ and can be manipulated as a standard algebraic symbol, e.g.

$$(1 - L)^2 y_t = (1 - 2L + L^2) y_t = y_t - 2y_{t-1} + y_{t-2}. \quad (1.15)$$

The MA(q) model may now be written as

$$y_t = \mu + \psi(L) \varepsilon_t,$$

where

$$\psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \dots + \psi_q L^q. \quad (1.16)$$

It was shown in section 1.3 that provided the conditions on the innovation series $\{\varepsilon_t\}$ hold, a process is stationary if the coefficients of the Wold representation satisfy

$\sum_{j=0}^{\infty} \psi_j^2 < \infty$. Since, for an MA(q) process, $\psi_j = 0 \forall j > q$, this summation is satisfied for all finite q and hence, provided the conditions on the innovation series $\{\varepsilon_t\}$ hold, an MA(q) process is stationary.

From equation 1.9, the ACF of an MA(q) process is given by

$$R(u) = \sigma^2 \sum_{j=0}^q \psi_j \psi_{j+u}, \quad (1.17)$$

where σ^2 is the variance of the innovations. Note, for $u > q$, this gives, by definition, $\psi_{j+u} = 0$ for $j \in (0, \dots, q)$ and thus the ACF vanishes after q lags.

Consider the general MA(q) process

$$y_t = \psi(L)\varepsilon_t, \quad (1.18)$$

where the mean may be taken as zero without loss of generality. The identity,

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i, \quad (1.19)$$

will be of use in the following discussion. Now, dividing equation 1.18 by $\psi(L)$ gives

$$\varepsilon_t = \frac{1}{\psi(L)} y_t \quad (1.20)$$

Factorising $\psi(L)$ into $(1 - \alpha_1 L)(1 - \alpha_2 L) \dots (1 - \alpha_q L)$, where the roots of the polynomial are $1/\alpha_i$ for $i \in (1, \dots, q)$, equation 1.20 can be written in the form

$$\varepsilon_t = \frac{A_1}{1 - \alpha_1 L} y_t + \frac{A_2}{1 - \alpha_2 L} y_t + \dots + \frac{A_q}{1 - \alpha_q L} y_t,$$

for constants $|A_1|, \dots, |A_q| < \infty$, and making use of identity 1.19, this can be written as

$$\varepsilon_t = A_1 \sum_{j=0}^{\infty} \alpha_1^j y_{t-j} + A_2 \sum_{j=0}^{\infty} \alpha_2^j y_{t-j} + \dots + A_q \sum_{j=0}^{\infty} \alpha_q^j y_{t-j}. \quad (1.21)$$

This is known as the *autoregressive*, AR, representation of the process. Although, as previously discussed, an MA(q) process with appropriate conditions on the innovations is stationary for any finite ψ , it can be seen that for each of these will converge if and only if $|\alpha_i| < 1$, for all $i \in (1, \dots, q)$, and hence the modulus of the roots of the polynomial $\psi(z)$ must all be greater than 1. If an MA process admits an AR representation, with AR coefficients, ϕ_j such that $\sum_{j=0}^{\infty} \phi_j^2 < \infty$, where

$$\phi_j = \sum_{i=1}^q A_i \alpha_i^j,$$

it is said to be *invertible*. Invertibility ensures the uniqueness of the ACF. Note, since $|\alpha_i| < 1$, for all $i \in (1, \dots, q)$, ϕ_j decay at an exponential rate. This exponential rate of decay implies the AR coefficients are absolutely summable, that is $\sum_{j=0}^{\infty} |\phi_j| < \infty$.

Considering the MA(1) process, substitution of the ACF into the Durbin-Levinson algorithm defined in equations 1.12, 1.13 and 1.14, gives the PACF to be of the form

$$\Phi(u) = -\psi^u \frac{(1 - \psi^2)}{1 - \psi^{2(u+1)}}.$$

Unlike the ACF which vanishes after lag q , the PACF thus remains for all u , decaying to zero at an exponential rate. Exact forms of the PACF for higher order MA(q) processes possess complicated forms, but it is widely known that provided the roots of $\psi(z)$ are all real and greater than 1 then the PACF will decay at an exponential rate, see for example Wei [1989].

An *autoregressive*, AR(p), process is a process $\{y_t\}$ such that y_t is a linear combination of $\{y_s\}$ with $s \in (t-1, t-2, \dots, t-p)$, plus noise, that is

$$(y_t - \mu) = \sum_{j=1}^p \phi_j (y_{t-j} - \mu) + \varepsilon_t$$

where μ is the mean of the process and $\{\varepsilon_t\}$ is white noise such that $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2 \forall t \in \mathbb{Z}_+$ and $E(\varepsilon_t \varepsilon_s) = 0$ for $s \neq t$. This can be written as,

$$\phi(L)(y_t - \mu) = \varepsilon_t$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \dots - \phi_p L^p. \quad (1.22)$$

For the general AR(p) process, similar arguments as those in equations 1.20-1.21 show for the process to be stationary, each of the modulus of the roots of the polynomial $\phi(z)$ must all be greater than 1 and a stationary AR(p) process can be written as

$$(y_t - \mu) = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where ψ_j decay at an exponential rate, that is $\psi_j \sim B\rho^j$ as $j \rightarrow \infty$, for some constant $|\rho| < 1$ and constant or slowly varying B .

Use of $\psi_j \sim B\rho^j$ gives

$$R(u) \sim \sigma^2 \sum_{j=0}^{\infty} B^2 \rho^{2j+u} = B^2 \rho^u \sigma^2 \sum_{j=0}^{\infty} \rho^{2j} = \frac{B^2 \sigma^2 \rho^u}{1 - \rho^2} = \tilde{B} \tilde{\rho}^u, \quad (1.23)$$

for some constant $|\tilde{\rho}| < 1$ and constant or slowly varying \tilde{B} . Hence the ACF of a stationary AR(p) process decays at an exponential rate.

Consider now the PACF of a general AR(p) process. From the definition of an AR(p) process,

$$\mathcal{P}(y_t | y_{t-1}, \dots, y_{t-u+1}) = \sum_{j=0}^p \phi_j y_{t-j} \quad \text{for all } u > p.$$

Notice that $y_t - \mathcal{P}(y_t | y_{t-1}, \dots, y_{t-u+1})$ is therefore given by ε_t , for all $u > p$, and that this gives the PACF as

$$\Phi(u) = \text{Corr}(\varepsilon_t, y_{t-u} - \mathcal{P}(y_{t-u} | y_{t-1}, \dots, y_{t-u+1})) = 0 \quad \forall u > p,$$

since ε_t is white noise and therefore uncorrelated with all past observations. Thus, similar to the ACF of an MA(q) process, the PACF of an AR(p) process vanishes after lag p .

Having considered AR(p) and MA(q) processes, the general *autoregressive moving average*, ARMA(p,q), is now introduced. An ARMA(p,q) process, $\{y_t\}$, is defined by as

$$\phi(L)(y_t - \mu) = \psi(L)\varepsilon_t \quad (1.24)$$

where $\phi(L)$ and $\psi(L)$ are defined by equation 1.22 and equation 1.16 respectively and $\{\varepsilon_t\}$ is white noise. For ease of notation, from this point on μ is assumed to be zero without loss of generality.

The MA representation of an ARMA(p,q) process is given by

$$y_t = \frac{\psi(L)}{\phi(L)} \varepsilon_t. \quad (1.25)$$

For uniqueness $\psi(L)$ and $\phi(L)$ are assumed to have no common roots. Similarly, the AR representation of this process is given by

$$\varepsilon_t = \frac{\phi(L)}{\psi(L)} y_t.$$

Arguing as before, for the general ARMA(p,q) process to be stationary each of the modulus of the roots of the polynomial $\phi(z)$ must all be greater than 1 and for the process to be invertible each of the modulus of the roots of the polynomial $\psi(z)$ must all be greater than 1.

Also with similar arguments as those presented for the AR(p) and MA(q) processes, the coefficients of the MA and AR representations of an invertible and stationary ARMA(p,q) process decay at an exponential rate. From equation 1.23, this implies the ACF of an ARMA(p,q) decays at an exponential rate. This rate of decay will be discussed further in Chapter 2.

1.6.1 Linear Filters

A linear filter, $\theta(L)$, is a filter such that

$$\theta(L) = \sum_{j=-\infty}^{\infty} \theta_j L^j, \quad (1.26)$$

where L is the lag function. The following theorem, widely available in the literature, see for example Brockwell and Davis [1991], Hannan [1980] and Rozanov [1967], will be used to give the effect of such a filter to an ARMA process.

Theorem 1.6.1. *Let $\{x_t\}$ be a zero mean stationary process and $\theta(L)$ be a linear filter as defined by equation 1.26, such that*

$$\sum_{j=-\infty}^{\infty} |\theta_j| < \infty$$

then

- a) The process $y_t = \theta(L)x_t$ is zero mean stationary
- b) If

$$x_t = \psi(L)\varepsilon_t,$$

where $\psi(L)$ is lag polynomial, then

$$y_t = \theta(L)\psi(L)\varepsilon_t = \psi(L)\theta(L)\varepsilon_t = \zeta(L)\varepsilon_t,$$

where $\zeta(L) = \psi(L)\theta(L)$.

For an ARMA(p, q) process, $\{x_t\}$, defined by equation 1.25, if $\theta(L)$ can be written as $\zeta(L)/\eta(L)$, where $\zeta(L)$ is a polynomial of finite order s and $\eta(L)$ is a polynomial of finite order h , then the process $\{y_t\}$ defined by

$$y_t = \theta(L)x_t = \frac{\zeta(L)}{\eta(L)} \frac{\psi(L)}{\phi(L)} \varepsilon_t$$

is an ARMA($p+h$, $q+s$) process. However, to ensure uniqueness, any common factors in the numerator and denominator must be cancelled out. Thus, if there are k such common factors, $0 \leq k \leq \min(p+h, q+s)$, the resulting process will be an ARMA($p+h-k$, $q+s-k$) process.

Define now the *autocovariance generating function* (ACGF), $g(L)$, of a process as

$$g(L) = \sum_{u=-\infty}^{\infty} R(u)L^u \quad (1.27)$$

such that the coefficient of L^u is the covariance at lag u . For any stationary process, $\{y_t\}$ with Wold representation, $y_t = \psi(L)\varepsilon_t$, where the order of $\psi(L)$ can be infinite, the ACGF is given by

$$\begin{aligned} g(L) &= \sum_{u=-\infty}^{\infty} (\sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i-u}) L^u \\ &= \sigma^2 \sum_{i=0}^{\infty} \psi_i L^i \sum_{h=0}^{\infty} \psi_h L^{-h} \\ &= \sigma^2 \psi(L^{-1}) \psi(L) \end{aligned} \quad (1.28)$$

For an ARMA process, with Wold representation given by

$$y_t - \mu = \frac{\psi(L)}{\phi(L)} \varepsilon_t$$

it follows from equation 1.28 that the ACGF of an ARMA model is given by

$$g(L) = \frac{\psi(L^{-1})\psi(L)}{\phi(L)^{-1}\phi(L)} \sigma^2 \quad (1.29)$$

Also, if two uncorrelated ARMA processes are added together, such that

$$y_t - \mu = \frac{\psi(L)}{\phi(L)} \varepsilon_t + \frac{\vartheta(L)}{\varphi(L)} \xi_t$$

where $E(\xi_t) = 0$, $E(\xi_t^2) = \sigma^2$, $E(\xi_t \xi_{t-u}) = 0$ for $u \neq 0$ and $E(\varepsilon_t \varepsilon_s) = 0$ for all s and t , the ACGF is given by

$$g(L) = \frac{|\psi(L)|^2}{|\phi(L)|^2} \sigma^2 + \frac{|\vartheta(L)|^2}{|\varphi(L)|^2} \sigma^2$$

since the cross product term will disappear when taking expectations. These properties can be extended to the multiplication of several Lag polynomials and the addition of three or more uncorrelated ARMA models.

1.6.2 Sample Mean and ACF

In this section the asymptotic distributions of the sample mean, sample covariance function and sample correlation function for ARMA processes are given. These distributions can then be used to carry out statistical inference for the observed time series. Reviews of asymptotic theory can be found by several authors, for example Serfling [1980] and Billingsley [1986].

The first theorem put forward will be the *Central Limit Theorem*, proving the normality of the sample mean of IID random variables. A proof can be found from, for example, Theorem 6.4.1 of Brockwell and Davis [1991].

Theorem 1.6.2. *Let $\{x_t\} \sim \text{IID}(\mu, \sigma^2)$ and \bar{x} be the sample mean, given by*

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t,$$

then

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2), \quad \text{as } n \rightarrow \infty$$

Now, consider an observed series, $\{y_t\}$, $t \in (1, \dots, n)$, taken from an ARMA(p,q) process. As discussed in section 1.6, the ACF of an ARMA(p,q) process decays at an exponential rate and is thus absolutely summable. The following theorem which can be found in Anderson [1971], amongst others, is therefore applicable and gives the asymptotic consistency of \bar{y} as $n \rightarrow \infty$.

Theorem 1.6.3. *Let $\{y_t\}$, $t \in (1, \dots, n)$ be an observed series from a linear stationary process with mean μ and ACF such that*

$$\sum_{u=-\infty}^{\infty} |R(u)| < \infty,$$

and

$$\sum_{u=-\infty}^{\infty} R(u) = C,$$

for some constant C . Let \bar{y} be the sample mean of $\{y_t\}$. Then, as $n \rightarrow \infty$, \bar{y} is \sqrt{n} consistent for μ .

Under the further assumption that the innovation series, $\{\varepsilon_t\}$, are IID with zero mean and finite variance, the asymptotic distribution of the sample mean for an ARMA(p,q) process is given by the following Theorem, taken from Theorem 8.4.1 of Anderson [1971].

Theorem 1.6.4. *Let $y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_t$, where $\{\varepsilon_t\}$, are IID with zero mean and finite variance, σ^2 , and $\sum_{j=-\infty}^{\infty} |\psi_j| = B < \infty$. Then, as $n \rightarrow \infty$,*

$$\sqrt{n}(\bar{y} - \mu) \rightarrow_d N(0, B^2 \sigma^2).$$

Note, Theorem 1.6.4 applies to ARMA(p,q) processes, which have been shown in section 1.6 to have exponentially decaying MA coefficients, and can also be applied to series which are not causal. This theorem does not, however, apply to series with ACF which are not absolutely sumable. The situation in which $\sum_{u=-\infty}^{\infty} r(u)$ does not converge is addressed in Chapter 2.

The next statistic to be examined will be an estimate of the autocovariance function, $R(u)$. The sample autocovariance of an observed time series $\{y_t\}$ at lag u is given by

$$\hat{R}(u) = \frac{1}{n} \sum_{t=u+1}^n (y_t - \bar{y})(y_{t-u} - \bar{y}) \quad (1.30)$$

with the sample variance given by $\hat{R}(0)$. The following theorem is due to Bartlett [1946].

Theorem 1.6.5. *Let $\{y_t\}$, $t \in (1, \dots, n)$, be an observed series from a stationary linear process with innovations $\{\varepsilon_t\}$ such that $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t \varepsilon_s) = 0$ when $t \neq s$, $E(\varepsilon_t^4) = \eta \sigma^4$ and $E(\varepsilon_t^2 \varepsilon_s^2) = E(\varepsilon_t^2)E(\varepsilon_s^2) = \sigma^4$ when $t \neq s$, $\forall t, s \in \mathbb{Z}$ and $\eta > 0$ and absolutely sumable ACF. Define $s(i, j)$ such that,*

$$s(i, j) = \sum_{k=-\infty}^{\infty} \left(R(k)R(k+i-j) + R(k+i)R(k-j) \right)$$

Then, the distribution of $\hat{R}(h) = (\hat{R}(1), \dots, \hat{R}(h))'$ as $n \rightarrow \infty$ is given by

$$\sqrt{n} \left(\hat{R}(h) - R(h) \right) \xrightarrow{d} N(0, \Sigma),$$

where $R(h) = (R(1), \dots, R(h))'$ and Σ is the matrix defined $\Sigma_{i,j} = s(i, j)$

The sample correlation at lag u is given by the sample covariance at lag u divided by the sample variance,

$$\hat{r}(u) = \frac{\hat{R}(u)}{\hat{R}(0)}. \quad (1.31)$$

The asymptotic properties of the sample autocorrelations were also studied by Bartlett [1946] and following theorem is due to his work.

Theorem 1.6.6. Let $\{y_t\}$, $t \in (1, \dots, n)$, be an observed series from a stationary linear process with innovations $\{\varepsilon_t\}$ such that $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, $E(\varepsilon_t \varepsilon_s) = 0$ when $t \neq s$, $E(\varepsilon_t^4) = \eta\sigma^4$ and $E(\varepsilon_t^2 \varepsilon_s^2) = E(\varepsilon_t^2)E(\varepsilon_s^2) = \sigma^4$ when $t \neq s$, $\forall t, s \in \mathbb{Z}$ and $\eta > 0$ and absolutely sumable ACF. Define $s(i, j)$ such that,

$$s(i, j) = \sum_{j=-\infty}^{\infty} \left(r(j)r(j+\nu-u) + r(j+\nu)r(j-u) + 2r(u)r(\nu)r(j)^2 \right. \\ \left. - 2r(u)r(j)r(j-\nu) - 2r(\nu)r(j)r(j-u) \right)$$

Then, the distribution of $\hat{\mathbf{r}}(h) = (\hat{r}(1), \dots, \hat{r}(h))'$ as $n \rightarrow \infty$ is given by

$$\sqrt{n}(\hat{\mathbf{r}}(h) - \mathbf{r}(h)) \xrightarrow{d} N(0, \Sigma),$$

where $\mathbf{r}(h) = (r(1), \dots, r(h))'$ and Σ is the matrix defined $\Sigma_{i,j} = s(i, j)$

Since the ACF of an ARMA(p,q) process decay at an exponential rate, Theorems 1.6.5 and 1.6.6 are applicable, under the additional assumptions on the innovations, to the sample autocovariances and autocorrelations.

1.6.3 Parameter Estimation

For an observed time series, $\{y_t\}$, for $t \in (1, \dots, n)$ assumed to be generated from an ARMA(p, q) process in a practical application, the parameters σ^2 , $\{\phi_i\}$ and $\{\psi_j\}$, for $i \in (1, \dots, p)$ and $j \in (1, \dots, q)$, will typically be unknown and must therefore be estimated from the observed series. The question of parameter estimation has been addressed by many authors, see for example Box and Jenkins [1976], Brockwell and Davis [1991] and Wei [1989] for reviews.

Yule-Walker Estimates

The set of *Yule-Walker* equations were introduced by Yule [1927] and Walker [1931]. Use of these equations to find estimates of the unknown parameters is more commonly applied to AR(p) processes rather than general ARMA(p,q) processes. The case in which $q = 0$ is thus considered first.

Let $\{y_t\}$ be an AR(p) process, therefore

$$\begin{aligned} r(u) &= \frac{1}{R(0)} \left(\sum_{j=0}^p \phi_j E((y_{t-j} - \mu)(y_{t-u} - \mu)) + E(\varepsilon_t(y_{t-u} - \mu)) \right) \\ &= \sum_{j=0}^p \phi_j r(u-j), \quad \text{for } u \neq 0. \end{aligned}$$

and

$$r(0) = \sum_{j=0}^p \phi_j r(j) + \sigma^2 = 1 \quad (1.32)$$

The set of equations 1.32 are known as the *Yule-Walker* equations, see Yule [1927] and Walker [1931]. These can be written as

$$\begin{pmatrix} 1 & r(1) & \dots & r(p-1) \\ r(1) & 1 & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & 1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{pmatrix}$$

thus

$$\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} 1 & r(1) & \dots & r(p-1) \\ r(1) & 1 & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{pmatrix}.$$

and

$$\sigma^2 = 1 - \sum_{j=0}^p \phi_j r(j)$$

This can be written in matrix form as

$$\Phi_p = \Gamma_p^{-1} \mathbf{r}_p,$$

and

$$\sigma^2 = 1 - \Phi_p' \mathbf{r}_p,$$

where $\Phi_p = (\phi_1, \phi_2, \dots, \phi_p)'$, $\Gamma_p = (r(|i-j|))_{i,j=1}^p$, and $\mathbf{r}_p = (r(1), r(2), \dots, r(p))'$. The matrix, Γ_p , is always invertible since, from Theorem 1.5.1, the ACF is positive-definite for any stationary process. If the true ACF is known, this gives exact values for the parameters. However, usually the ACF is not known. Estimates of the parameters can be found by replacing the autocorrelations with their sample estimates, $\hat{r}(1), \hat{r}(2), \dots, \hat{r}(p)$, found using equations 1.30 and 1.31. Thus, the Yule Walker estimates are given by

$$\hat{\Phi}_p = \hat{\Gamma}_p^{-1} \hat{r}_p, \quad (1.33)$$

and

$$\hat{\sigma}^2 = 1 - \hat{\Phi}_p' \hat{r}_p, \quad (1.34)$$

The following theorem, taken from Theorem 8.1.1 of Brockwell and Davis [1991], gives the asymptotic behaviour of the Yule-Walker estimates. The result is originally due to Mann and Wald [1943].

Theorem 1.6.7. *Let $\{y_t\}$ be an $AR(p)$ process, with $\{\varepsilon_t\} \sim IID(0, \sigma^2)$ and let $\hat{\Phi}_p$ and $\hat{\sigma}^2$ be the Yule-Walker estimates of Φ and σ^2 defined by equation 1.33 and equation 1.34 respectively. Then*

$$\sqrt{n}(\hat{\Phi}_p - \Phi_p) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Gamma_p^{-1})$$

and

$$\hat{\sigma}^2 \rightarrow \sigma^2$$

as $n \rightarrow \infty$.

Theorem 1.6.7 gives the asymptotic distribution of the parameters when a true order, p , is assumed to exist and be known. Theorem 8.1.2 of Brockwell and Davis [1991] also gives the following result regarding the asymptotic behaviour of the Yule-Walker estimates if the order p is misspecified.

Theorem 1.6.8. *Let $\{y_t\}$ be an $AR(p)$ process, with $\{\varepsilon_t\} \sim IID(0, \sigma^2)$, let m be the misspecified order $m > p$ and let $\hat{\Phi}_m$ be the Yule-Walker estimates of Φ_m defined by equation 1.33, where Φ_m is the coefficient vector of the best linear predictor $\Phi_m' y_m$ of y_{m+1} . Then*

$$\sqrt{n}(\hat{\Phi}_m - \Phi_m) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \Gamma_m^{-1})$$

as $n \rightarrow \infty$. In particular

$$\sqrt{n}(\hat{\phi}_m) \xrightarrow{d} N(0, 1)$$

as $n \rightarrow \infty$.

Theorem 1.6.8 will be particularly useful in the order selection of an $AR(p)$ process. This will be discussed further in section 1.6.4.

In order to avoid the inversion of the matrix $\hat{\Gamma}_p$ when finding the Yule-Walker estimates, see equation 1.33, it is possible to use the Durbin-Levinson Algorithm,

see section 1.5, substituting $\hat{r}(u)$ for $r(u)$. Durbin [1960] and Levinson [1947] show for an AR(p) process, the estimated values of $\phi_{p1}, \phi_{p2}, \dots, \phi_{pp}$ are identical to the Yule-Walker estimates of ϕ_1, \dots, ϕ_p , but the procedure is computationally faster. In addition to avoiding the matrix inversion, this algorithm also provide estimates of the PACF, $\hat{\phi}(u)$. These estimates are useful for determining the order of an AR(p) process and will be discussed further in section 1.6.4.

The Yule-Walker equations can be adjusted for an ARMA(p, q) model. However, when $q > 0$, the corresponding equations are non-linear in the unknown parameters. This leads to possible non-existence and non-uniqueness of solutions and requires an iterative approach. As such, the Yule-Walker estimates for ARMA(p,q) models with $q > 0$ are often not very efficient, see, for example, Shumway and Stoffer [2000].

Maximum Likelihood Estimates

The *maximum likelihood* estimates (MLEs) of the parameters are the estimates which, for a given series $\{y_t\}$ and assumed model, i.e. ARMA(p,q), maximise the likelihood function

$$L(\mathbf{y}; \phi) = \prod_{t=1}^n f(y_t | Y_{t-1})$$

where $Y_{t-1} = \{y_s\}_{s < t}$ contains all the previous observations and f is probability density function based on the fitted model, with parameters ϕ . Maximising the log of this function is equivalent, but generally preferred since the log-likelihood function is a linear function given by

$$l(\phi) = \log(L(\mathbf{y}; \phi)) = \sum_{t=1}^n \log(f(y_t | Y_{t-1})).$$

Assuming $\{\varepsilon_t\}$ is Gaussian white noise, i.e. $\varepsilon_t \sim \text{IID } N(0, \sigma^2)$, $\forall t \in \mathbb{Z}$, the conditional distribution of $\{y_t\}$ for an AR(p) process is IID normally distributed, with

$$E(y_t | Y_{t-1}) = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p}, \quad \text{Var}(y_t | Y_{t-1}) = \sigma^2, \quad \forall t \in \mathbb{Z}.$$

From the normal distribution, this gives

$$f(y_t | Y_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\phi(L)y_t)^2}{2\sigma^2}\right).$$

It can then be seen that

$$l(\phi, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |\mathbf{V}_p| - \frac{1}{2\sigma^2} \left(\mathbf{y}'_p \mathbf{V}_p^{-1} \mathbf{y}_p + \sum_{t=p+1}^n (\phi(L)y_t)^2 \right) \quad (1.35)$$

where $\phi = (\phi_1, \dots, \phi_p)$ is the parameter vector and $\sigma^2 \mathbf{V}_p$ is the theoretical covariance matrix of $\mathbf{y}_p = (y_1, \dots, y_p)^n$ for the fitted model in terms of ϕ , which must be included since the observations y_{-p+1}, \dots, y_0 have not been observed. Taking the partial derivative with respect to σ^2 and equating to zero gives the MLE estimate for σ^2 as

$$\hat{\sigma}^2 = \frac{\left(\mathbf{y}'_p \mathbf{V}_p^{-1} \mathbf{y}_p + \sum_{t=p+1}^n (\phi(L)y_t)^2 \right)}{n} = \frac{S(\phi)}{n}$$

and substituting this back into equation 1.35 gives

$$l(\phi, \sigma^2) = -\frac{n}{2} \log(S(\phi)) - \frac{1}{2} \log |\mathbf{V}_p| + C(n)$$

where, for a fixed n , $C(n)$ is a constant and thus does not effect the maximisation procedure.

For the full ARMA(p, q) model, under the assumption that $\{\varepsilon_t\}$ are IID $N(0, \sigma^2)$, the conditional distribution of y_t is still normal with

$$E(y_t | Y_{t-1}) = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \psi_1 \varepsilon_{t-1} + \dots + \psi_q \varepsilon_{t-q}, \quad \text{Var}(y_t | Y_{t-1}) = \sigma^2.$$

Notice that the values of $\{\varepsilon_t\}$ have not been observed and must therefore be estimated. One method of estimating the innovation series, $\{\varepsilon_t\}$, is via the following recursive equations, known as the *Innovations Algorithm*, see for example, Brockwell and Davis [1996]

$$\begin{aligned} \hat{y}_1 &= 0 \\ \hat{y}_{t+1} &= \sum_{j=1}^t c_{tj}(y_{t+1-j} - \hat{y}_{t+1-j}) & 1 \leq t < \max(p, q) \\ \hat{y}_{t+1} &= \sum_{j=1}^p \phi_j y_{t+1-j} + \sum_{j=1}^q \psi_j \hat{e}_{t+1-j} & t \geq \max(p, q) \end{aligned}$$

$$\begin{aligned} \tau_0 &= R(0) \\ c_{t,t-k} &= \tau_k^{-1} (R(t-k) - \sum_{j=0}^{k-1} c_{k,k-j} c_{t,t-j} \tau_j), \quad k = 0, 1, \dots, t-1, \\ \tau_t &= R(0) - \sum_{j=0}^{t-1} c_{t,t-j}^2 \tau_j, \end{aligned}$$

where

$$\tau_t = E((y_t - \hat{y}_t)^2)$$

and

$$\hat{e}_t = y_t - \hat{y}_t.$$

The innovations algorithm finds the coefficients c_{qi} in the MA(q) model $y_t = \sum_{i=0}^q c_{qi} \varepsilon_{t-i}$ and is thus comparable to the Durbin-Levinson algorithm which finds the coefficients ϕ_{pi} in the AR(p) model $y_t = \varepsilon_t + \sum_{i=1}^p \phi_{pi} y_{t-i}$. An alternative method of finding estimates of the innovations, suggested by Hannan and Rissanen [1982], is to estimate y_t from fitting a high order AR model using the Durbin-Levinson algorithm and take $\hat{e}_t = y_t - \hat{y}_t$. Note, the innovations algorithm assumes the unknown past values of $\{y_t\}$ and $\{\varepsilon_t\}$ for $t \leq 0$ are all zero. Box and Jenkins [1976] suggests instead use of back forecasts to estimate the unknown past observations from the observed $\{y_t\}$. As the series length n increases the methods become asymptotically equivalent.

The log-likelihood function is given by

$$l(\phi, \psi) = -\frac{n}{2} \log(S(\phi, \psi)) - \frac{1}{2} \log |V_n| + C(n)$$

where $\psi = (\psi_1, \dots, \psi_q)'$ is the vector of MA parameters,

$$\frac{S(\phi, \psi)}{n} = \frac{\mathbf{y}'_n \mathbf{V}_n^{-1} \mathbf{y}_n}{n} = \hat{\sigma}^2$$

with $\mathbf{y}_n = (y_1, \dots, y_n)'$ and $\sigma^2 \mathbf{V}_n = E(\mathbf{y}_n \mathbf{y}'_n)$ is the theoretical covariance matrix of the fitted model in terms of ϕ and ψ .

The calculation and inversion of \mathbf{V}_n can be avoided by noticing, see Brockwell and Davis [1991], that $\sigma^2 \mathbf{V}_n = \mathbf{C} \mathbf{D} \mathbf{C}'$, where \mathbf{C} is the $n \times n$ lower triangular matrix defined by

$$\mathbf{C} = [c_{i,i-j}]_{i,j=0}^{n-1}$$

and $D = \text{diag}(\tau_0, \dots, \tau_{n-1})$, where c_{jk} and τ_j are defined by the innovations algorithm. Now, the innovations algorithm can be expressed in matrix form as

$$\hat{\mathbf{y}}_n = (\mathbf{C} - \mathbf{I})(\mathbf{y}_n - \hat{\mathbf{y}}_n)$$

thus

$$\mathbf{y}_n = \mathbf{C}(\mathbf{y}_n - \hat{\mathbf{y}}_n).$$

This gives

$$\mathbf{y}'_n \mathbf{V}_n^{-1} \mathbf{y}_n = (\mathbf{y}_n - \hat{\mathbf{y}}_n)' \sigma^2 \mathbf{D}^{-1} (\mathbf{y}_n - \hat{\mathbf{y}}_n) = \sigma^2 \sum_{j=1}^n \frac{(y_j - \hat{y}_j)}{\tau_{j-1}}$$

and

$$|\mathbf{V}_n| = \frac{|\mathbf{C}|^2 |\mathbf{D}|}{\sigma^2} = \frac{|\mathbf{D}|}{\sigma^2} = \prod_{j=0}^{n-1} \frac{\tau_j}{\sigma^2}.$$

The MLEs are then given by the parameters which minimise the function

$$l(\phi, \psi) = \log(S(\phi, \psi)) + \frac{1}{n} \sum_{j=0}^{n-1} \log(\tau_j / \sigma^2) \quad (1.36)$$

where

$$S(\phi, \psi) = \sigma^2 \sum_{j=1}^n \frac{(y_j - \hat{y}_j)^2}{\tau_{j-1}} = (n - q - p) \hat{\sigma}^2.$$

The asymptotic distribution of the MLEs is given in the following theorem, the proof of which can be found in Brockwell and Davis [1991].

Theorem 1.6.9. *Let $\{y_t\}$ be an ARMA(p, q) process in reduced form defined by equation 1.24, and let $\hat{\phi}$ and $\hat{\psi}$ be the estimates of ϕ and ψ which maximise equation 1.36, then*

$$\frac{S(\hat{\phi}, \hat{\psi})}{n - p - q} \xrightarrow{d} \frac{\sigma^2 \chi_{n-p-q}^2}{n - p - q}$$

$$\sqrt{n}((\hat{\phi}', \hat{\psi}')' - (\phi', \psi')') \xrightarrow{d} N(0, \sigma^2 \Sigma),$$

where, for $p, q \geq 1$,

$$\Sigma = \begin{pmatrix} E(\mathbf{U}_t \mathbf{U}_t') & E(\mathbf{U}_t \mathbf{W}_t') \\ E(\mathbf{W}_t \mathbf{U}_t') & E(\mathbf{W}_t \mathbf{W}_t') \end{pmatrix}^{-1}$$

with $U_t = (u_t, \dots, u_{t+1-p})'$ defined by

$$\phi(L)u_t = \varepsilon_t$$

and $W_t = (w_t, \dots, w_{t+1-q})'$ defined by

$$\psi(L)w_t = \varepsilon_t.$$

For $p = 0$, $\Sigma = E(W_t W_t')^{-1}$ and for $q = 0$, $\Sigma = E(U_t U_t')^{-1}$.

The maximum likelihood procedure can be simplified by noticing that since $\tau_j \rightarrow \sigma^2$ as $j, n \rightarrow \infty$ for all stationary invertible ARMA(p, q) processes, then the second term in equation 1.36 is asymptotically negligible. Thus the minimisation of equation 1.36 is asymptotically equivalent to minimising

$$S(\phi, \psi) \rightarrow \sum_{j=1}^n (y_j - \hat{y}_j)^2. \quad (1.37)$$

The estimates of ϕ and ψ based on this procedure are asymptotically equivalent for IID Gaussian innovations to the *Least Squares*, LS, estimates, presented below.

Least Squares Estimates

Let $\{y_t\}$ be an observed series of length n taken from an ARMA(p, q) process with innovations $\{\varepsilon_t\}$ which are $IID(0, \sigma^2)$. If the parameters and past values of $\{e_t\}$ are known, the best linear predictor of y_t based on past values $\{y_s\}$ and $\{\varepsilon_s\}$ for $s < t$ is given by

$$\hat{y}_t = \sum_{j=1}^q \psi_j \varepsilon_{t-j} + \sum_{j=1}^p \phi_j y_{t-j},$$

and thus $\varepsilon_t = y_t - \hat{y}_t$, can be considered the error term. When the $\{\varepsilon_t\}$ and the parameters are unknown, the $\{\varepsilon_t\}$ can be estimated using methods such as the innovations algorithm. The *least squares* method of parameter estimation then estimates the parameters by choosing the values which minimise the sum of squares of the error terms, that is

$$\hat{\psi}, \hat{\phi} = \arg \min \sum_{j=1}^n (y_j - \hat{y}_j)^2. \quad (1.38)$$

Comparison of equation 1.38 with equation 1.37 shows for IID Gaussian innovations, the least squares estimates are asymptotically equivalent to the maximum likelihood estimates. The asymptotic distribution of the LS estimates, for a series with Gaussian innovations, are therefore also given by Theorem 1.6.9.

1.6.4 Model Selection

For an observed time series, $\{y_t\}$, assumed to be taken from an ARMA(p, q) process, the estimates of the ARMA parameters discussed in section 1.6.3 depend on the knowledge of the order p and q . When these orders are unknown they must be estimated from the available $\{y_t\}$.

The classical approach to model selection for an ARMA(p, q) process, see Box and Jenkins [1976], is graphical in nature, using plots of the sample ACF and sample PACF. As discussed in section 1.6, for an MA(q) model, from equation 1.17, $R(u) \neq 0$ when $u \leq q$ and $R(u) = 0$ when $u > q$ and as such the sample ACF would be expected to be significantly different from zero only for the first q lags. Tests can be carried out using the asymptotic distributions given by Theorem 1.6.6. For an AR(p) process, the ACF will decay at an exponential rate, but the PACF will disappear after lag p . Anderson [1942] shows the sample PACF are approximately $N(0, 1/n)$ for large n . For an ARMA(p, q) process, both the ACF and PACF decay at an exponential rate, however, the exponential rate of decay for the PACF starts after lag p and the exponential rate of decay of the ACF starts after lag q , thus inspection of the sample ACF and sample PACF may still provide clues of the actual order.

After studying the sample ACF and sample PACF to gain initial estimates of p and q and fitting the parameters using methods such as those discussed in section 1.6.3, Box and Jenkins [1976] then suggests model diagnostic checking be carried out to test the goodness of fit and, if the model proves inadequate, discover in what way it is inadequate. These results are then used to suggest a new model and this new model is then tested. This is repeated until a suitable model is found.

Let $\{\hat{\varepsilon}_t\}$ be the series of residuals for the fitted model, that is $\hat{\varepsilon}_t = y_t - \hat{y}_t$, where \hat{y}_t is the value of y_t suggested by the model. Let $r_\varepsilon(u)$ be the sample ACF at lag u for this residual series. Box and Pierce [1970] derive the variances of these residuals and show for large n and large u , $\text{Var}(r_\varepsilon(u)) \approx 1/n$, whilst for small u the variances can be substantially less than this. These variances can be used to test if the residuals are uncorrelated and thus if the fitted ARMA(p, q) model is adequate in capturing the time dependence of the observed series.

An alternative to testing each of the $r_\varepsilon(u)$ is to note, from Theorem 1.6.6, that since the sample ACF of an uncorrelated series is asymptotically normally distributed with variance approximately equal to $1/n$ for large n , then it is possible to show, see Box and Pierce [1970], that

$$n \sum_{k=1}^K r_\varepsilon(k)^2 \rightarrow_d \chi^2(K - p - q)$$

and hence a Portmanteau goodness of fit test can be carried out.

If the model proves to be inadequate, that is the residuals appear correlated, Box

and Jenkins [1976] suggest fitting a model to the residuals and substituting this back into the original.

The model selection procedure described by Box and Jenkins [1976] is quite strongly subjective to personal choice. Motivated by the desire to remove this element of personal choice, Akaike [1974] proposed the use of an information criterion.

Let $\{x_t\}$ be a series of random variables of length n taken from a density $g(x)$. If two models are considered, with probability density functions $f_1(\cdot|\theta_1)$ and $f_2(\cdot|\theta_2)$, with parameter vectors θ_1 with dimension p_1 and θ_2 with dimension p_2 , such that $p_1 > p_2$, and the maximum likelihood estimates of θ_1 and θ_2 , then it is known, see for example Huber [1967], that, as $n \rightarrow \infty$,

$$D = 2(l_1(\hat{\theta}_1) - l_2(\hat{\theta}_2)) \rightarrow_d \chi_{p_1-p_2}^2, \quad (1.39)$$

where l_1 and l_2 are the log-likelihood functions based on f_1 and f_2 . For choosing between two models a hypothesis test can then be carried out, which rejects f_2 in favour of f_1 if $D > \chi_{(p_1-p_2),\alpha}^2$ at a significance level of $100\alpha\%$. For choosing between many models such tests become complicated.

Now, from equation 1.39, the asymptotic expectation of D is $2(p_1 - p_2)$. Consider the case in which f_1 is favoured to f_2 if $D > 2(p_1 - p_2)$, that is, if D is larger than expected. This is asymptotically equivalent to

$$2(l_1(\hat{\theta}_1) - l_2(\hat{\theta}_2)) > 2(p_1 - p_2)$$

$$2(l_1(\hat{\theta}_1) - p_1) > 2(l_2(\hat{\theta}_2) - p_2)$$

$$-2(l_1(\hat{\theta}_1) - p_1) < -2(l_2(\hat{\theta}_2) - p_2).$$

Extending this to choosing between multiple models suggests choosing the model which minimises $-2(l_1(\hat{\theta}) - p)$. For ARMA(p,q) models, this suggests the use of selecting the order of p and q by minimising the criterion

$$\text{AIC}(p, q) = -2 \log L(\hat{\theta}_{p,q}) + 2(p + q),$$

where $L(\hat{\theta}_{p,q})$ is the maximised likelihood function. Although Akaike [1974] originally suggested the name AIC simply to mean the first such information criterion, which could then be followed by BIC, DIC, IIC etc, AIC is often referred to as 'Akaike's information criterion'. Work by Shibata [1976], however, proves the use of AIC does not produce consistent estimates of p and q . The following result, valid for AR(p) processes, is based on Theorem 1 of Shibata [1976].

Theorem 1.6.10. *Let $\{y_t\}$ be an observed series of length n taken from an AR(p_0) process, with true order p_0 . Let \hat{p} be the order selected using AIC to fit an AR model*

to $\{y_t\}$ over the range $p \in (0, \dots, P)$, for some $P \geq p_0$. Let $\{x_t\}$ be a series of IID random variables with distribution χ_1^2 and let $S_j = \sum_{t=1}^j x_t$. Then, as $n \rightarrow \infty$,

$$P(\hat{p} = p) \rightarrow \begin{cases} a_{(p-p_0)} b_{(P-p)} & p_0 \leq p \leq P \\ 0 & \text{otherwise} \end{cases}$$

where

$$a_j = P(S_1 > 2, S_2 > 4, \dots, S_j > 2^j)$$

$$b_j = P(S_1 < 2, S_2 < 4, \dots, S_j < 2^j)$$

and $a_0, b_0 = 1$.

Theorem 1.6.10 shows that although the probability of underestimating the true order for fitting an AR model when using AIC tends to zero, the probability of overestimating the order tends to a nonzero constant and thus the estimate is not consistent. Hannan [1980] extended this result, proving that use of AIC also has a tendency to overestimate the order of q for an MA(q) process and suggested a similar result would hold for the general ARMA(p, q) process. Similar criteria have been introduced which alter the AIC in order to remove this inconsistency.

Bhansali and Downham [1977] introduced a generalised version of AIC of the form

$$AIC_\alpha(p, q) = -2 \log L(\theta_{p,q}) + \alpha(p + q),$$

for some constant α , arguing that although the choice of $\alpha = 2$ was justified by Akaike [1974] using information-theoretic considerations, the use of $\alpha \neq 2$ should also be considered. They showed that the asymptotic probability of choosing the correct model using AIC_α increased with α and suggested use of $\alpha > 1$.

Akaike [1977], Rissanen [1978] and Schwarz [1978] considered the use of the criterion

$$BIC(p, q) = -2 \log L(\theta_{p,q}) + \log(n)(p + q),$$

and Hannan and Quinn [1979] introduced the criterion

$$HIC(p, q) = -2 \log L(\theta_{p,q}) + 2c \log(\log(n))(p + q),$$

for some $c > 1$. The asymptotic properties of the estimated orders for fitting ARMA(p, q) models to a given series generated from an ARMA(p_0, q_0) process using these criteria were studied by Hannan [1980], in which the following result is proved.

Theorem 1.6.11. *Let $\{y_t\}$ be an observed series of length n taken from an ARMA(p_0, q_0) process, with true order p_0, q_0 and innovations such that $E(\varepsilon_t | \{\varepsilon_s : s < t\}) = 0$, $E(\varepsilon_t^2 | \{\varepsilon_s : s < t\}) = \sigma^2$ and $E(\varepsilon_t^\gamma) < \infty$, $\gamma \geq 4$, for all $t \in (1, \dots, n)$. Let p_b and q_b be the estimates of p_0 and q_0 found using BIC from the range $p, q \in (0, \dots, K)$ for $p_0, q_0 \leq K$. Then, as $n \rightarrow \infty$,*

$$p_b \rightarrow_p p_0$$

and

$$q_b \rightarrow_p q_0$$

If in addition the innovations are independent and p_h and q_h are the estimates of p_0 and q_0 obtained from use of HIC, then, as $n \rightarrow \infty$,

$$p_h \rightarrow_p p_0$$

and

$$q_h \rightarrow_p q_0$$

In addition to showing the consistency of BIC and HIC, Hannan [1980] also proves the consistency of the generalised criterion

$$AIC_\alpha(p, q) = -2 \log L(\theta_{p,q}) + \alpha(p + q),$$

where α is no longer considered to be constant, but instead $\alpha \rightarrow \infty$ as $n \rightarrow \infty$. This result agrees with the previously mentioned result of Bhansali and Downham [1977].

1.7 The Frequency Domain

The study of the frequency domain and the spectral density for empirical functions began with Stokes [1879] and Schuster [1899], who developed the use of the modulus-squared Fourier transform to search for hidden periodicities. Slutsky [1929], [1934] and Wiener [1930] went on to study many of the statistical properties of the spectral density and propose a general form of harmonic analysis for general stationary processes. Since then, frequency analysis has been used in many fields such as physics, acoustics, economics, biology and psychology, amongst many others.

1.7.1 Fourier Analysis

Let $\{y_t\}$ be a series of length n . The Fourier representation of $\{y_t\}$ is given by writing $\{y_t\}$ as a linear combination of trigonometric terms. First, define λ_j as

$$\lambda_j = \frac{2\pi j}{n} \quad \text{for } j \in (1, 2, \dots, [n/2]),$$

where $[n/2]$ is the integer part of $n/2$. These λ_j are known as the Fourier frequencies. Now, the Fourier series of $\{y_t\}$ is given by

$$y_t = \sum_{j=0}^{[n/2]} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)), \quad (1.40)$$

where a_j and b_j are the Fourier coefficients, where

$$b_j = \frac{2}{n} \sum_{t=1}^n y_t \sin(\lambda_j t) \quad j \in (1, \dots, [(n-1)/2])$$

and

$$a_j = \frac{2}{n} \sum_{t=1}^n y_t \cos(\lambda_j t) \quad j \in (1, \dots, [(n-1)/2])$$

$$a_0 = \frac{1}{n} \sum_{t=1}^n y_t$$

$$a_{[n/2]} = \frac{1}{n} \sum_{t=1}^n y_t (-1)^t \quad (\text{when } n \text{ is even}).$$

due to the orthogonality relationships between \cos and \sin ,

$$\sum_{t=1}^n \sin\left(\frac{2\pi kt}{n}\right) \sin\left(\frac{2\pi jt}{n}\right) = \begin{cases} n/2, & k = j \neq 0, n/2 \\ 0, & \text{otherwise} \end{cases}$$

$$\sum_{t=1}^n \cos\left(\frac{2\pi kt}{n}\right) \cos\left(\frac{2\pi jt}{n}\right) = \begin{cases} n/2, & k = j \neq 0, n/2 \\ n, & k = j = 0, n/2 \\ 0, & k \neq j \end{cases}$$

and

$$\sum_{t=1}^n \cos\left(\frac{2\pi kt}{n}\right) \sin\left(\frac{2\pi jt}{n}\right) = 0,$$

where k and j are integers in the range from 0 to $n/2$.

Since $\exp(i\lambda) = \cos(\lambda) + i\sin(\lambda)$, the Fourier series of $\{y_t\}$ can also be written as

$$y_t = \sum_{j=-[(n-1)/2]}^{[n/2]} c_j e^{i\lambda_j t},$$

where

$$c_0 = a_0$$

$$c_j = \frac{a_j - ib_j}{2}$$

$$c_{-j} = \frac{a_j + ib_j}{2},$$

thus

$$c_j = \frac{1}{n} \sum_{t=1}^n y_t e^{-i\lambda_j t}. \quad (1.41)$$

Transforming a series $\{y_t\}$ into the series $\{c_j\}$ as shown in equation 1.41 is related to the discrete Fourier transform, $w(\lambda)$, where

$$w(\lambda) = \frac{1}{2\pi} \sum_{t=1}^n y_t e^{-i\lambda t}.$$

It can immediately be seen that

$$c_j = \frac{2\pi}{n} w(\lambda_j),$$

and hence

$$y_t = \sum_{j=-[(n-1)/2]}^{[n/2]} \frac{2\pi}{n} w(\lambda_j) e^{i\lambda_j t}.$$

It can be seen that no information is lost during the transformation and that any bounded series of finite length n can be represented in this way. Now, note that $\Delta\lambda = 2\pi/n$ is equal to $\lambda_j - \lambda_{j-1}$ and consider the case when $n \rightarrow \infty$,

$$\begin{aligned}
 y_t &= \lim_{n \rightarrow \infty} \sum_{j=-\lfloor (n-1)/2 \rfloor}^{\lfloor n/2 \rfloor} w(\lambda_j) e^{i\lambda_j t} \Delta\lambda & (1.42) \\
 &= \lim_{\Delta \rightarrow 0} \sum_{j=-\infty}^{\infty} w(\lambda_j) e^{i\lambda_j t} \Delta\lambda \\
 &= \int_{-\pi}^{\pi} w(\lambda) e^{i\lambda t} d\lambda.
 \end{aligned}$$

The transformation of $w(\lambda)$ back to $\{y_t\}$ shown in equation 1.42 is known as the inverse discrete Fourier transform. In order for this representation to exist $w(\lambda)$ must be integrable. It is known that for this to hold the condition $\sum_{-\infty}^{\infty} y_t^2 < \infty$ must be satisfied, see for example Wei [1989].

1.7.2 Spectral Density

The spectral density for a stationary time series, $\{y_t\}$, is defined by the Fourier transform of the autocovariances

$$f(\lambda) = \frac{1}{2\pi} \left(\sum_{u=-\infty}^{\infty} R(u) e^{-i\lambda u} \right). \quad (1.43)$$

where $\lambda \in [-\pi, \pi]$ represents the frequency in radians. The spectral density can be standardised by dividing throughout by the variance and leaving it in terms of the correlations. The spectral density at $\lambda = 0$ is given by

$$f(0) = \frac{1}{2\pi} \left(\sum_{u=-\infty}^{\infty} R(u) \right).$$

Note for processes which have an absolutely summable ACF, for example ARMA(p, q) processes, $f(0)$ is a finite constant. The requirement for $f(\lambda)$ to be integrable however is the weaker condition that $\sum_{u=-\infty}^{\infty} R(u)^2 < \infty$ and thus the spectral density can have a singularity at zero yet still be integrable. Since for a stationary time series this square summable condition on the ACF holds, the spectral density is integrable and the inverse Fourier transform gives

$$R(u) = \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda u} d\lambda. \quad (1.44)$$

In particular, the variance of $\{y_t\}$ is defined by

$$R(0) = \int_{-\pi}^{\pi} f(\lambda) d\lambda.$$

If the function $f(\lambda)$ defined by equation 1.43 is not integrable, the variance of $\{y_t\}$ is thus undefined and hence $\{y_t\}$ is not stationary. Note, from equations 1.43 and 1.44 it can be seen that for a stationary time series the ACF and spectral density contain the same information and thus the time series can be defined by either. The following result is based on Corollary 4.1.1 of Brockwell and Davis [1996].

Theorem 1.7.1. *A square sumable function $R(u)$ is the autocovariance function of a stationary time series if and only if it is even and*

$$f(\lambda) = \frac{1}{2\pi} \left(\sum_{u=-\infty}^{\infty} R(u) e^{-i\lambda u} \right) \geq 0, \quad \text{for all } \lambda \in (-\pi, \pi]$$

in which case $f(\lambda)$ is the spectral density.

The nonnegative property of the spectral density shown in Theorem 1.7.1 is clear from the nonnegative-definite property of the ACF given in Theorem 1.5.1.

From the definition of $f(\lambda)$ given in equation 1.43, if $\{y_t\} \in \mathbb{R}$, then $f(\lambda)$ can be equivalently written as

$$f(\lambda) = \frac{1}{2\pi} \left(\sum_{u=-\infty}^{\infty} R(u) \cos(\lambda u) \right),$$

and hence $f(\lambda)$ is even and periodic with period 2π .

For a completely uncorrelated random white noise time series, ε_t , as defined in Section 1.3, the covariances are $R(u) = 0$ when $u \neq 0$ and $R(0) = \sigma^2$. Hence

$$f_{\varepsilon}(\lambda) = \frac{1}{2\pi} \left(\sigma^2 + 2 \sum_{u=1}^{\infty} (0) \cos(\lambda u) \right) = \frac{\sigma^2}{2\pi}$$

This gives a completely flat spectrum. A smooth series, with mostly positive ACF, will produce a large value of $f(0)$ and thus have more weight at lower frequencies. A jagged series, with more negative values of $r(u)$ will produce a smaller value of $f(0)$ and larger values at $f(\pi)$ and thus have more weight at higher frequencies.

A peak in the spectral density at $\lambda = \lambda^*$ shows a possible sign of periodicity of period $2\pi/\lambda^*$. The detection of hidden periodicity in this way was the original reason the spectral density was considered, see Stokes [1879] and Schuster [1899].

1.7.3 Periodogram

For a series $\{y_t\}$ of length n , section 1.7.1 showed the y_t could be written in terms of the Fourier series

$$y_t = \sum_{j=-(n-1)/2}^{[n/2]} c_j e^{i\lambda_j t},$$

where the Fourier coefficients $\{c_j\}$ give the weights of the relative Fourier frequencies. A measure of the distribution of $\{y_t\}$ over the various Fourier frequencies is given by the periodogram, $I(\lambda_j)$, defined as

$$I(\lambda_j) = \frac{n}{2\pi} |c_j|^2 \quad (1.45)$$

$$= \frac{1}{2\pi} \left(\sum_{u=-(n-1)}^{n-1} \hat{R}(u) e^{-i\lambda_j u} \right),$$

where $\hat{R}(u)$ is the sample autocovariance at lag u . The periodogram was first introduced by Schuster [1899]. Comparison of equation 1.45 with equation 1.43 shows the periodogram acts as an estimate of the spectral density at the Fourier frequencies. Consider the case in which the ACF of $\{y_t\}$ is absolutely summable. Since $\hat{R}(u)$ is \sqrt{n} -consistent for $R(u)$, see Theorem 1.6.5, the periodogram, $I(\lambda_j)$, will also be asymptotically unbiased for $f(\lambda_j)$.

The following theorem gives the sampling distribution of the periodogram from a process with absolutely summable ACF, see for example Wei [1989].

Theorem 1.7.2. *Let $\{y_t\}$ be a series of length n taken from a stationary process with absolutely summable ACF. Let $I(\lambda_j)$ be the periodogram defined by equation 1.45. Then, for $j, k \in (1, \dots, [n/2])$,*

$$I(\lambda_j) \sim f(\lambda_j) \frac{\chi_2^2}{2}$$

and

$$\text{Cov}(I(\lambda_j), I(\lambda_k)) = 0, \quad \text{for } j \neq k.$$

Note, Theorem 1.7.2 gives $E(I(\lambda_j)) = f(\lambda_j)$ and $\text{Var}(I(\lambda_j)) = f(\lambda_j)^2$. The variance does therefore not depend on n and does not decrease to zero as $n \rightarrow \infty$. Although $I(\lambda_j)$ is unbiased, it is not consistent for $f(\lambda_j)$, and since the $I(\lambda_j)$ are independent this results in often very jagged estimates of the spectral density. The

condition that the ACF are absolutely sumable is assumed for the following methods of producing consistent estimates of $f(\lambda_j)$.

One method of reducing the variance of the periodogram is that of smoothing using weighted averaging over $2m_n$ consecutive values, where m_n is allowed to vary with n such that

$$\hat{f}(\lambda_j) = \sum_{i=-m_n}^{m_n} \theta_i I(\lambda_{j-i}),$$

where

$$\sum_{i=-m_n}^{m_n} \theta_i = 1$$

$$\theta_i = \theta_{-i}$$

and

$$\sum_{i=-m_n}^{m_n} \theta_i^2 \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Under the assumption that $f(\lambda_j)$ is approximately constant over $[\lambda_{j-m_n}, \lambda_{j+m_n}]$, the mean and variance of $\hat{f}(\lambda_j)$ are given by

$$E(\hat{f}(\lambda_j)) = \sum_{i=-m_n}^{m_n} \theta_i E(I(\lambda_{j-i})) \quad (1.46)$$

$$\approx f(\lambda_j) \sum_{i=-m_n}^{m_n} \theta_i = f(\lambda_j)$$

and

$$\text{Var}(\hat{f}(\lambda_j)) = \sum_{i=-m_n}^{m_n} \theta_i^2 f^2(\lambda_{j-i}) \quad (1.47)$$

$$\approx f(\lambda_j)^2 \sum_{i=-m_n}^{m_n} \theta_i^2 \rightarrow 0, \quad \text{as } n \rightarrow \infty,$$

and therefore $\hat{f}(\lambda_j)$ is consistent for $f(\lambda_j)$. Due to this smoothing, neighbouring values of the spectrum will no longer be independent. The amount of dependency between $\hat{f}(\lambda_j)$ and $\hat{f}(\lambda_k)$ will be proportional to the amount of overlap of in the

windows used to estimate them. Equations 1.46 and 1.47 can thus be used again to show $\hat{f}(\lambda_j)$ is consistent for $f(\lambda_j)$.

Another method of estimating the spectral density is to first fit an AR(p) model to the data, using methods such as those given in sections 1.6.3 and 1.6.4. The theoretical spectrum for this fitted AR(p) model then becomes the estimator, $f_{AR}(\lambda)$, for the spectrum, $f(\lambda)$. This method is known as AR spectral estimation. It was suggested by Akaike [1969] and Parzen [1974]. As $n \rightarrow \infty$, Parzen [1974] proved the variance of $f_{AR}(\lambda)$ to be of the form

$$\frac{n}{2p} \text{Var}(f_{AR}(\lambda)) \rightarrow f(\lambda)^2.$$

Berk [1974] showed that, under the assumption that $p \rightarrow \infty$ as $n \rightarrow \infty$ at a sufficiently slow rate and the innovations are IID($0, \sigma^2$) with $E(\varepsilon_t^4) < \infty$, the AR spectral method is asymptotically equivalent to the method of smoothing. Hence $f_{AR}(\lambda)$ is consistent for $f(\lambda)$.

Bhansali [1997] extends the result of Berk [1974] to include the case when the innovations are α -stable. The following result is based on Theorem 3.1 of Bhansali [1997].

Theorem 1.7.3. *Let $\{y_t\}$ be a series of length n taken from a process*

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where

$$E(\varepsilon_t) = 0, P(|\varepsilon_t| > x) = x^{-\alpha} L(x),$$

for some slowly varying function $L(x)$ and $0 < \alpha < 2$ and

$$P(\varepsilon_t > x)/P(|\varepsilon_t| > x) \rightarrow 1/2, P(\varepsilon_t < -x)/P(|\varepsilon_t| > x) \rightarrow 1/2,$$

and

$$\sum_{j=0}^{\infty} |j| |\psi_j|^\delta < \infty,$$

where $\delta = \min(1, \alpha)$. Let $f_{AR}(\lambda)$ be the AR spectral estimate of $f(\lambda)$ based on the fitted model AR(p) and let $\nu > \delta$ and $0 < \beta < \min(0.5, \alpha^{-1} - \nu^{-1})$. Then, as $n \rightarrow \infty, p \rightarrow \infty, p/n^\beta \rightarrow 0$,

$$n^{1/\nu} \sup_{\lambda \in [-\pi, \pi]} |f_{AR}(\lambda) - f(\lambda)| \rightarrow_p 0.$$

A related estimate of the spectral density is known as the ARMA spectral estimation, see Gray and Woodward [1986]. As suggested by the name, this involves the fitting of an ARMA(p, q) to the data and taking the estimate of the spectral density to be the theoretical spectral density of the fitted model. As with the AR spectral estimate the quality of the ARMA spectral estimate is dependent on appropriate choices for p and q .

1.7.4 SGF, Gain and Phase

By comparing equation 1.43 with equation 1.27 a connection can be seen between the spectral density and the ACGF. Replacing L with $e^{-i\lambda}$ gives a spectral generating function, SGF, where

$$f(\lambda) = \frac{g(e^{-i\lambda})}{2\pi}.$$

The properties of the ACGF given in section 1.6.1 can thus be exploited to find the spectral density. For example, the effect of applying a linear filter $W(L)$ to a process $\{x_t\}$ with spectral density $f_x(\lambda)$, i.e.

$$y_t = W(L)x_t,$$

making use of the results of section 1.6.1 gives

$$f_y(\lambda) = |W(e^{-i\lambda})|^2 f_x(\lambda), \quad (1.48)$$

where $f_y(\lambda)$ is the spectral density of $\{y_t\}$. The function $W(e^{-i\lambda})$ is known as the frequency response function, and $|W(e^{-i\lambda})|^2$ is called the power transfer function. Note $W(e^{-i\lambda})$ is actually the Fourier transform of the coefficients of the linear filter,

$$W(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} W_j e^{-ij\lambda}$$

Now, consider the form of $\{x_t\}$ given by equation 1.42

$$x_t = \int_{-\pi}^{\pi} w(\lambda) e^{i\lambda t} d\lambda,$$

where $w(\lambda)$ is the Fourier transform of $\{x_t\}$. The effect of applying a linear filter to $\{x_t\}$ is the same as applying the Fourier transform of the filter to $w(\lambda)$. Therefore

$$y_t = W(L)x_t = \int_{-\pi}^{\pi} W(e^{-i\lambda}) w(\lambda) e^{i\lambda t} d\lambda.$$

Now, write $W(e^{-i\lambda}) = W_1(\lambda) + iW_2(\lambda)$, where $W_1(\lambda)$ and $W_2(\lambda)$ are both real, thus, in polar form,

$$W(e^{-i\lambda}) = |W(e^{-i\lambda})| e^{i \tan^{-1} \left(\frac{W_2(\lambda)}{W_1(\lambda)} \right)}. \quad (1.49)$$

The factor by which the amplitude of $w(\lambda)$ is enhanced or diminished by the linear filter is called the gain, $G(\lambda)$. From equation 1.49, this is given by the modulus of the frequency response function

$$G(\lambda) = |W(e^{-i\lambda})|$$

The shift in $w(\lambda)$ caused by the filter is called the phase, $Ph(\lambda)$, where

$$Ph(\lambda) = \tan^{-1} \left(\frac{-W_2(\lambda)}{W_1(\lambda)} \right).$$

This is derived from

$$\begin{aligned} y_t = W(L)x_t &= \int_{-\pi}^{\pi} |W(e^{-i\lambda})| e^{i \tan^{-1} \left(\frac{W_2(\lambda)}{W_1(\lambda)} \right)} w(\lambda) e^{i\lambda t} d\lambda \\ &= \int_{-\pi}^{\pi} G(\lambda) w(\lambda) e^{i\lambda t - iPh(\lambda)} d\lambda \\ &= \int_{-\pi}^{\pi} G \left(\lambda + \frac{Ph(\lambda)}{t} \right) w \left(\lambda + \frac{Ph(\lambda)}{t} \right) e^{i\lambda t} d\lambda, \end{aligned}$$

where the limits on the integral need not be changed since the function is periodic with period 2π . The shift in the time domain can also be seen by noting

$$y_t = \int_{-\pi}^{\pi} G(\lambda) w(\lambda) e^{i\lambda(t - Ph(\lambda)/\lambda)} d\lambda,$$

and hence the change in t caused by the linear filter $W(L)$ is given by $Ph(\lambda)/\lambda$. This shift will typically be depended on λ .

1.8 Exp Model

The *Exponential model*, $\text{Exp}(s)$, was introduced by Bloomfield [1973] as an alternative to the $\text{ARMA}(p,q)$ model. Instead of fitting a model to the series in the time domain, the $\text{Exp}(s)$ model specifies a form on the spectral density. An $\text{Exp}(s)$ process, $\{y_t\}$, has spectral density $f(\lambda)$ of the form

$$f(\lambda) = \frac{\sigma^2}{2\pi} \exp \left(2 \sum_{j=1}^s c_j \cos(j\lambda) \right).$$

From the discussion concerning the effect of linear filters given in section 1.7.4, it can be seen that

$$f(\lambda) = \frac{\sigma^2}{2\pi} \exp \left(2 \sum_{j=1}^s c_j \cos(j\lambda) \right) = \frac{\sigma^2}{2\pi} |\psi(e^{-i\lambda})|^2 = \frac{\sigma^2}{2\pi} |\phi(e^{-i\lambda})|^{-2}, \quad (1.50)$$

where $\psi(L)$ and $\phi(L)$ are the infinite order lag polynomials such that

$$y_t = \psi(L)\varepsilon_t \quad \text{and} \quad \phi(L)y_t = \varepsilon_t,$$

where $\{\varepsilon_t\}$ is an uncorrelated mean zero series with variance $\sigma^2 < \infty$. Equation 1.50 thus gives

$$\psi(L) = \exp \left(\sum_{j=1}^s c_j L^j \right) \quad (1.51)$$

and

$$\phi(L) = \exp \left(\sum_{j=1}^s -c_j L^j \right). \quad (1.52)$$

Consider the Exp(1) process. From the Taylor expansion

$$\psi(L) = \sum_{j=1}^{\infty} \frac{c_1^j}{j!} L^j \quad (1.53)$$

and

$$\phi(L) = \sum_{j=1}^{\infty} \frac{(-c_1)^j}{j!} L^j. \quad (1.54)$$

Hence the MA and AR coefficients $\{\psi_j\}$ and $\{\phi_j\}$ eventually decay at an exponential rate as $j \rightarrow \infty$, provided c_1 is not unbounded. For $|c_1| < 1$ this is immediately clear from equations 1.53 and 1.54. For $|c_1| \geq 1$, note that

$$\psi_j = \frac{c_1}{j} \psi_{j-1}$$

and for bounded c_1 , $c_1/j \rightarrow 0$ as $j \rightarrow \infty$. Similar arguments hold for $\{\phi_j\}$. Thus the Exp(1) process is stationary and invertible with exponentially decaying MA and AR coefficients. The ACF of the Exp(1) process, from equation 1.9, is given by

$$R(u) = \sigma^2 \sum_{j=1}^{\infty} \frac{c_1^{2j+u}}{j!(j+u)!},$$

which also decays at an exponential rate as $u \rightarrow \infty$. For the general Exp(s) process, the Taylor expansions of equations 1.51 and 1.52 give

$$\psi(L) = \prod_{k=1}^s \left(\sum_{j=1}^{\infty} \frac{c_k^j}{j!} L^j \right) \quad (1.55)$$

and

$$\phi(L) = \prod_{k=1}^s \left(\sum_{j=1}^{\infty} \frac{(-c_k)^j}{j!} L^j \right).$$

Similar arguments to those for the Exp(1) process show the Exp(s) process with finite s is stationary and invertible with MA and AR coefficients, and ACF, that decay at an exponential rate. All of the previous theorems stated for the ARMA(p, q) process which require only conditions on the innovations and absolutely sumable ACF will therefore still hold for a Exp(s) process under the same conditions on the innovations. In particular, the sample mean \bar{y} is \sqrt{n} -consistent for the mean μ by Theorem 1.6.3.

When the order $s \rightarrow \infty$, conditions must then be placed on the coefficients c_j to ensure stationarity. From equation 1.55, the coefficient ψ_1 of the lag polynomial $\psi(L)$ is given by $\sum_{j=1}^{\infty} c_j$ and the coefficients c_j should therefore be absolutely sumable. In addition, for a finite Exp(s) model to approximate the infinite order Exp(∞) process, the coefficients must tend to zero.

Methods of estimating the parameters of an Exp(s) model and generalisations of the model are discussed in Chapter 2.

2. LONG-MEMORY PROCESSES

2.1 Definition of a Long-Memory Process

In order to understand what is meant by 'long-memory' it is first useful to consider what is meant by 'short-memory', so as to contrast the two. Let $\{y_t\}$ be a stationary invertible ARMA(p, q) process with finite p and q . It is shown in section 1.6 that $\{y_t\}$ can be written with an infinite AR or MA representation with absolutely summable coefficients which decrease at an exponential rate. It is also shown that the autocorrelations, $r(u)$, also decay at an exponential rate as $u \rightarrow \infty$, that is

$$r(u) \sim B\rho^{|u|} \quad 0 < \rho < 1, \quad \text{as } u \rightarrow \infty \quad (2.1)$$

As mentioned in section 1.6, this exponential decay rate implies that the $r(u)$ are absolutely summable. In addition, this absolute summability of the ACF shows Theorem 1.6.3 is applicable and the sample mean is \sqrt{n} consistent.

Note, since the spectral density function is defined as

$$f(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} R(u) \exp(-iu\lambda),$$

the absolute summability of the ACF, along with the positive-definite property of the ACF, see Theorem 1.5.1, ensures that $0 < f(\lambda) < \infty$ for all $\lambda \in [0, \pi]$. In particular, when $\lambda = 0$,

$$f(0) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} R(u) = C < \infty, \quad (2.2)$$

for some constant C . A stationary process with the properties given in equation 2.1 and equation 2.2 is said to be a short-memory process.

However, it was observed that certain stationary discrete-time series displayed an ACF which decreased to zero slower than an exponential rate. This was found to be the case, for example, in certain data sets from hydrology and climatology, see Hurst [1951],[1957], Mandelbrot and Wallis [1968], Mandelbrot [1972] and M'Leod and Hipel [1978] amongst others. The absolute summation of the ACFs of these time series did not appear to converge to a finite number. Due to this, M'Leod and

Hipel [1978] defined a long-memory process as a stationary process with ACF, $r(u)$, such that

$$\lim_{n \rightarrow \infty} \sum_{u=0}^n |r(u)| \rightarrow \infty. \quad (2.3)$$

From equation 2.2 it can be seen that this is true if and only if the spectral density is unbounded when $\lambda = 0$. It can also be seen that Theorem 1.6.3 is no longer applicable and the sample mean may not be consistent.

One method of classifying long-memory is to assume that the autocorrelations of a process $\{y_t\}$ decrease at a polynomial rate with,

$$r(u) \sim B(u)u^{2d-1}, \quad \text{as } u \rightarrow \infty \quad (2.4)$$

where $B(u)$ is a constant or slowly varying function as $u \rightarrow \infty$ and d is referred to as the *memory parameter*. For $d \in (0, 0.5)$, it is widely known that the ACF of $\{y_t\}$ satisfies equation 2.3 and thus has long-memory. It is important to note, however, that equation 2.3 does not imply equation 2.4 and thus the two definitions are not equivalent.

Another widely used definition for long memory is that the spectral density of the process is of the form

$$f(\lambda) \sim G(1/\lambda)\lambda^{-2d} \quad \text{as } \lambda \rightarrow 0+ \quad (2.5)$$

with $G(1/\lambda)$ being a constant or slowly varying function as $\lambda \rightarrow 0$. Note, for $d \in (0, 0.5)$ this is unbounded and thus the process has long memory. For $d = 0$, equation 2.5 tends to a constant if $G(1/\lambda) \rightarrow G_0$, with $0 < G_0 < \infty$, as $\lambda \rightarrow 0+$, thus the process will have short memory. If $d \in (-0.5, 0)$, equation 2.5 tends to zero as $\lambda \rightarrow 0+$. This is known as an *intermediate memory process*, which is known to occur, for example, when over differencing takes place, see section 2.2.2.

A third often used definition for long memory comes from taking the coefficients of $\{y_T\}$ in the Wold Decomposition, see Theorem 1.3.1, as

$$\psi_j \sim j^{d-1}l(j), \quad \text{as } j \rightarrow \infty, \quad (2.6)$$

where $l(j)$ is a constant or slowly varying function as $j \rightarrow \infty$. The connections between equations 2.4, 2.5 and 2.6 are given in the following Theorem, a proof of which can be found in Palma [2007].

Theorem 2.1.1. *Let $\{y_t\}$ be a stationary regular process. Take $d \in (0, 0.5)$. Then,*

- a) *if $\{y_t\}$ satisfies equation 2.6 then it also satisfies equation 2.4*
- and, *if $B(u)$ is quasi-monotone slowly varying then*
- b) *if $\{y_t\}$ satisfies equation 2.4 then it also satisfies equation 2.5.*

Despite increasing popularity in the literature there is no universally agreed definition of long memory. Although equations 2.4-2.6 give the more classical definitions, these may not be suitable for non-stationary time series, see Heyde and Yang [1997] and Hall [1996]. One almost equivalent alternative to equation 2.3 is by looking at the variance of the sample mean directly, that is, let

$$y^m = \frac{\sum_{t=1}^m y_t}{m}$$

and investigate

$$\lim_{m \rightarrow \infty} m \text{Var}(y^m).$$

If this limit tends to a constant, the sample mean is \sqrt{m} consistent and the process is said to have short memory, whilst if it is unbounded the process is said to have long memory, see Heyde and Yang [1997].

In the review of long memory by Guegan [2005], a process $\{y_t\}$ with spectral density $f(\lambda)$ is considered to possess long memory if

$$\frac{\sup_{\lambda \in [-\pi, \pi]} f(\lambda)}{\inf_{\lambda \in [-\pi, \pi]} f(\lambda)}$$

is unbounded. This definition includes the case when the spectral density is unbounded at $\lambda = 0$, though also allows for situations when the spectral density is unbounded at points away from $\lambda = 0$. A similar definition also given in Guegan [2005], states the process has long memory if, for $d \in (0, 0.5)$ and some $\lambda_0 \in [-\pi, \pi]$

$$f(\lambda) \sim G(1/\lambda) |\lambda|^{-2d}, \quad \text{as } \lambda \rightarrow \lambda_0,$$

where $G(1/\lambda)$ is a constant or slowly varying function as $\lambda \rightarrow \lambda_0$.

To avoid confusion, for the remainder of this thesis a process is said to have long memory if it is stationary and its ACF is not absolutely sumable. Any further assumptions on the process will be stated. The case in which the singularity of the spectral density is away from zero will not be considered further in this thesis, a review of such processes can be found in Palma [2007].

2.2 Long-Memory Models

2.2.1 Fractional Gaussian Noise

Brownian motion, $B(t)$, $t \in \mathbb{R}$, is a continuous Gaussian stochastic process such that $B(0) = 0$ almost surely, $B(t)$ has independent increments, the expected value of $B(t)$ is independent of t and $\text{Var}(B(t) - B(s)) = \sigma^2 |t - s|$ for some constant σ , $s \in \mathbb{R}$.

A standard *fractional Brownian motion*, $B_d(t)$, is defined as

$$B_d(t) = \frac{\sqrt{\Gamma(2d+2) \cos(\pi d)}}{\Gamma(d+1)} \int_{-\infty}^{\infty} (t-s)_+^d - (-s)_+^d dB(s) \quad (2.7)$$

where

$$s_+ = \begin{cases} s & \text{for } s \geq 0 \\ 0 & \text{for } s < 0, \end{cases}$$

and $\Gamma(x)$ is the gamma function. The following theorem gives properties of the mean, variance, covariance and distribution of $B_d(t)$.

Theorem 2.2.1. *Let $B_d(t)$ be standard fractional Brownian motion as defined in equation 2.7, then, for $d \in (-0.5, 0.5)$ and $t, s \in \mathbb{R}$,*

$$E(B_d(t)) = 0,$$

$$\text{Cov}(B_d(t), B_d(s)) = \frac{1}{2} (|t|^{2d+1} + |s|^{2d+1} - |t-s|^{2d+1}).$$

$$B_d(t) \sim N(0, |t|^{2d+1})$$

Now, let $\{y_t : t \in \mathbb{Z}\}$ be the sequence defined by

$$y_t = B_d(t+1) - B_d(t), \quad (2.8)$$

then y_t is known as standard *fractional Gaussian noise* (fGN). The following theorem, see Tawfik [2003], gives some of the properties of $\{y_t\}$.

Theorem 2.2.2. *Let $\{y_t : t \in \mathbb{Z}\}$ be standard fractional Gaussian noise, as defined by equation 2.8. Then*

- a) $\{y_t\}$ is stationary for $d \in (-0.5, 0.5)$
- b) $y_t \sim N(0, 1)$
- c) The ACF of $\{y_t\}$ is given by

$$r(u) = \frac{1}{2} (|u+1|^{2d+1} + |u-1|^{2d+1} - 2|u|^{2d+1}),$$

which, for $d \neq 0$, is of the form

$$r(u) \sim d(2d+1) |u|^{2d-1}, \quad \text{as } u \rightarrow \infty.$$

- d) The spectral density of $\{y_t\}$ is given by

$$f_y(\lambda) \sim \left(\frac{2\Gamma(2d+2)}{\pi} \right) \sin(\pi(d+0.5))\lambda^{-2d}$$

The polynomial rate of decay of the ACF for implies the ACF is not absolutely summable for $d > 0$ and thus Theorem 1.6.3 regarding the consistency of the sample mean does not apply. Note, from the definition of $\{y_t\}$ given in equation 2.8,

$$\bar{y} = \frac{1}{n} (B_d(n+1) - B_d(1)),$$

thus, from Theorem 2.2.1,

$$\bar{y} \sim N(0, n^{2d-1}),$$

for all $n \in \mathbb{N}$.

The class of *general* fGN processes, see Geweke and Porter-Hudak [1983], are those with spectral density of the form $f(\lambda) = f_y(\lambda)f_x(y)$, where $f_y(\lambda)$ is the spectral density of fGN and $f_x(\lambda)$ is the spectral density of a short memory process. One such class of models is the set of ARMA models with fGN as the series of innovations. Since $f_x(\lambda) \rightarrow C$ as $\lambda \rightarrow 0$, for positive constant C , the rate at which $f(\lambda) \rightarrow \infty$ for general fGN processes are the same as that for fGN.

2.2.2 FARIMA model

Let $\{y_t\}$ be a non-stationary time series, for example, a random walk

$$y_t = y_{t-1} + \varepsilon_t = \sum_{j=-\infty}^t \varepsilon_j,$$

where $\{\varepsilon_t\}$ are IID with mean zero and variance σ^2 , say. Although the original series, $\{y_t\}$, is non-stationary with infinite variance, $\sum_{-\infty}^t \sigma^2$, taking the increments of such a series, $\{y_t - y_{t-1}\}$, can produce a stationary series, in this case $\{\varepsilon_t\}$. The ARIMA(p,d,q) model was designed to handle such series. It specifies that the observed time series, $\{y_t\}$, has the property

$$(1 - L)^d y_t = x_t \tag{2.9}$$

where $d \in \mathbb{Z}$ is the number of times the series should be differenced, L is the Lag function such that $Ly_t = y_{t-1}$ and $\{x_t\}$ follows an ARMA(p,q) model. The random walk mentioned is thus an ARIMA(0,1,0) process.

Granger and Joyeux [1980] noted however that ARIMA processes could cause over differencing, and that econometricians had been reluctant to use such models

for fear of 'zapping out the low frequency components'. This referred to the situation in which although the original time series was found to be non-stationary with infinite variance, the spectral density of the differenced series was zero at the origin. On considering this difficulty, Granger and Joyeux [1980] introduced the *fractional ARIMA*, FARIMA(p,d,q), model as an extension of the ARIMA(p,d,q) in which the parameter d in equation 2.9 could take non-integer values. Granger and Joyeux [1980] showed for $d > 0.5$, the series is non-stationary. For $d \in (0, 0.5)$ the series is a stationary long memory process, whilst for $d \in (-0.5, 0)$ the series has intermediate memory. Other early works on fractional differencing include Granger [1980], [1981], and Hosking [1981].

It is a well known result that, for $d < 0.5, d \neq 0, -1, -2, \dots$,

$$(1 - L)^d = \sum_{j=0}^{\infty} b_j L^j = b(L)$$

with

$$b_j = \frac{\Gamma(j - d)}{\Gamma(j + 1)\Gamma(-d)}$$

where $\Gamma(x)$ is the gamma function. This implies for $d \notin \mathbb{Z}$ the polynomial $b(L)$ is invertible. Another well known result is that

$$\sum_{j=0}^{\infty} b_j^2 < \infty,$$

for $d < 0.5$. Hence, a FARIMA(p,d,q) process with $d < 0.5$ is stationary and invertible if the short memory component $\{x_t\}$ is stationary and invertible, the conditions for which are given in section 1.6. The following theorem, see Kokoszka and Taqqu [1995], gives the form of the infinite MA and AR expansions of a stationary invertible FARIMA process and details of the ACF.

Theorem 2.2.3. *Let y_t be a FARIMA(p,d,q) process as defined by equation 2.9. Then*

$$y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$$

where $\theta_0 = 1$ and, for $d \in (-1, 0) \cup (0, 0.5)$,

$$\theta_j \sim B \frac{j^{d-1}}{\Gamma(d)}, \quad \text{as } j \rightarrow \infty$$

and

$$y_t = \sum_{j=1}^{\infty} \phi_j y_{t-j} + \varepsilon_t$$

where for $d \in (-1, 0) \cup (0, 0.5)$,

$$\phi_j \sim B \frac{j^{-d-1}}{\Gamma(-d)}, \quad \text{as } j \rightarrow \infty$$

for some B such that $0 < |B| < \infty$.

Also, if the innovations ε_t have mean zero and variance σ^2 . Then, for $d \in (-1, 0) \cup (0, 0.5)$, as $u \rightarrow \infty$

$$R(u) \sim B \frac{\sigma^2}{\pi} \Gamma(1-2d) \sin(\pi d) u^{2d-1},$$

for some positive constant B .

Note, when $d = 0$, $\{y_t\}$ is an ARMA(p,q) process and thus has correlations which decay at an exponential rate. Granger and Joyeux [1980] point out that an alternative view of $d = 0$ in terms of the ACF would produce correlations which decay at a harmonic rate,

$$r(u) \sim Bu^{-1},$$

for some positive constant or slowly varying B . The FARIMA model can not produce such correlations. A new extension to the FARIMA model is introduced in Chapter 5 which can produce such correlations.

This next theorem, see for example Palma [2007], gives the rate of convergence for the sample mean of a FARIMA(p,d,q) process.

Theorem 2.2.4. *Let $\{y_t\}$ $t \in (1, \dots, n)$ be a sample from a stationary invertible FARIMA(p,d,q) process as defined by equation 2.9, with innovations ε_t with mean zero and variance σ^2 . Let \bar{y} be the sample mean. Then, for $d \in (0, 0.5)$, as $n \rightarrow \infty$*

$$\text{Var}(\bar{y}) \sim B \frac{\sigma^2}{\pi} \frac{\Gamma(1-2d) \sin(\pi d)}{d(2d+1)} n^{2d-1}$$

for some positive constant B .

The following theorem gives the form of the spectral density of a FARIMA process. It follows directly from equation 1.48 and the identity $\lim_{x \rightarrow 0} \sin(x)/x \rightarrow 1$.

Theorem 2.2.5. Let $\{y_t\}$ be a FARIMA(p, d, q) process, defined by equation 2.9 with $\{x_t\}$ being an ARMA(p, q) process. Let $f_y(\lambda)$ and $f_x(\lambda)$ be the spectral densities of $\{y_t\}$ and $\{x_t\}$ respectively. Then

$$\begin{aligned} f_y(\lambda) &= |1 - e^{-i\lambda}|^{-2d} f_x(\lambda) \\ &= (2 \sin^2 \frac{\lambda}{2})^{-d} f_x(\lambda), \end{aligned}$$

which, as $\lambda \rightarrow 0$, gives

$$f_y(\lambda) \sim f_x(0) |\lambda|^{-2d}.$$

Comparison of the rate of decay of correlations for a FARIMA process and the rate at which the spectral density of a FARIMA process tends to infinity as $\lambda \rightarrow 0$ with the corresponding results for fGN given in Theorem 2.2.2 show them to be the same in terms of d . Theorem 1 of Geweke and Porter-Hudak [1983] considers this relationship and proves a process $\{y_t\}$ is a FARIMA(p, d, q) process if and only if it is also a general fGN process with parameter d .

2.2.3 FExp model

The fractional exponential, FExp(s), model extends the Exp set of models introduced by Bloomfield [1973]. Early work on FExp models is due to Beran [1993] and Robinson [1994a]. Similar to the FARIMA set of models, the FExp model also assumes y_t to be in the form given in equation 2.9, where $\{x_t\}$ is now a Exp process instead of an ARMA process. The spectral density, $f_y(\lambda)$, of a FExp process, $\{y_t\}$ is of the form

$$f_y(\lambda) = (2 \sin^2 \frac{\lambda}{2})^{-d} f_x(\lambda), \quad (2.10)$$

where $f_x(\lambda)$ is the spectral density of $\{x_t\}$, which, see section 1.8, is of the form

$$\log(f_x(\lambda)) = \sum_{j=0}^s c_j \cos(j\lambda), \quad (2.11)$$

for $c_0, \dots, c_s \in \mathbb{R}$. Substitution of equation 2.11 in equation 2.10 gives

$$\log(f_y(\lambda)) = -d \log(2 \sin^2 \frac{\lambda}{2}) + \sum_{j=0}^s c_j \cos(j\lambda).$$

Section 1.8 shows the ACF of Exp processes have an exponential rate of decay, as do the coefficients in the infinite MA and AR expansions. The spectral density for an Exp process is a bounded and positive for all $\lambda \in [-\pi, \pi]$. Therefore, from

the definition of general fGN it can be seen that an FExp(s) process is also a general fGN process and thus, by Theorem 1 of Geweke and Porter-Hudak [1983], also a FARIMA(p,d,q) process, although possibly with infinite p and q . The asymptotic results of Theorems 2.2.3, 2.2.4 and 2.2.5 regarding the infinite MA and AR expansion coefficients, the rate of decay of correlations, the rate of convergence of the sample mean and the behaviour of the spectral density near zero thus still hold for FExp processes. Indeed, these results hold for any short memory process x_t , see Granger and Joyeux [1980].

The FExp model can be generalized in various ways. Diggle [1990] gives an example which uses

$$f_x(\lambda) = \exp\left(\sum_{i=0}^s c_i |\lambda|^i\right),$$

for $c_0, \dots, c_s \in \mathbb{R}$.

A generalised definition of the model is given in Beran [1993]. It is said that $\{y_t\}$ is an FExp process with short memory components h_1, \dots, h_s and long memory component g if its spectral density is given by

$$f_y(\lambda) = g(\lambda)^{-2d} \exp\left(\sum_{i=0}^s c_i h_i(\lambda)\right)$$

Where $g : [-\pi, \pi] \rightarrow \mathbb{R}_+$ is an even positive function such that

$$\lim_{x \rightarrow 0} \frac{g(x)}{x} \rightarrow 1.$$

Also $h_0 = 1$ and $h_i, i \neq 0$ are even piecewise continuous functions over the range $[-\pi, \pi]$ where for any n , the $n^* \times (p+1)$ matrix with column vectors

$$\left(h_k\left(\frac{2\pi}{n}\right), h_k\left(\frac{2\pi 2}{n}\right), h_k\left(\frac{2\pi 3}{n}\right), \dots, h_k\left(\frac{2\pi n^*}{n}\right) \right)^T \quad \text{for } k = 0, \dots, s,$$

is non singular.

The conditions on the short memory components h_i ensure that $\exp\left(\sum_{i=0}^s c_i h_i(\lambda)\right)$ behaves like a spectral density of a short memory process, that is remain positive and bounded, whilst the conditions on $g(\lambda)$ give the shape of $f_y(\lambda)$ as $\lambda \rightarrow 0$ to still be approximately equal to that given in equation 2.5.

2.3 Estimation of the Long Memory Parameter

2.3.1 FARIMA Method

The FARIMA method of estimating the long memory parameter d is a *parametric* method. This means, for an observed series $\{y_t\}$, a fully parametrised model is fitted via the maximisation of a likelihood function, and estimates of the short and long range parameters are found simultaneously. The likelihood methods of estimation discussed in this section are not limited to FARIMA models and thus stated in a more general form. The results presented here are due to Fox and Taquq [1986], Dahlhaus [1989] and Giraitis and Surgailis [1990].

Let $\{y_t\}$ be a stationary Gaussian process sequence with spectral density $f_y(\lambda, \theta)$, and autocovariances, $R(u, \theta)$, where $\theta \in \Theta \in R^p$ are the p unknown parameters, with true parameters $\theta_0 \in \Theta$ and Θ is compact. If $\theta \neq \theta'$ the set $\{\lambda | f_y(\lambda, \theta) = f_y(\lambda, \theta')\}$ is supposed to have positive Lebesgue measure. The forms of $f_y(\lambda, \theta)$ and $R(u, \theta)$, in terms of λ , u and θ , are assumed known, as is the number of parameters required, p . Estimation of p when unknown will be discussed later.

For a Gaussian process, $\{y_t\}$, $t \in (1, \dots, n)$, the log-likelihood function is known to be

$$L(\theta) = -\frac{1}{2} (\log \det \Gamma_\theta + \mathbf{y}' \Gamma_\theta^{-1} \mathbf{y}), \quad (2.12)$$

where $\mathbf{y} = (y_1, \dots, y_n)'$, and Γ_θ is the covariance matrix of $\{y_t\}$ dependent on the unknown parameter vector, θ . The maximum likelihood (ML) estimate, $\hat{\theta}$, is the value of θ which maximises $L(\theta)$. Fox and Taquq [1986] and Dahlhaus [1989] give the asymptotic distribution of $\hat{\theta}$ under several conditions, omitted here, on the spectral density and its partial derivatives. These conditions are satisfied for FARIMA(p,d,q) processes, as well as fractional Gaussian noise, see Fox and Taquq [1986], and FExp(s) processes. The following theorem is due to Fox and Taquq [1986] and Dahlhaus [1989].

Theorem 2.3.1. *Let y_t be an observed time series of length n generated from a FARIMA(p,d,q) model with Gaussian innovations, true parameters $\theta_0 = (d, \theta_1, \dots, \theta_p, \phi_1, \dots, \phi_q)'$, spectral density $f_y(\lambda, \theta_0)$, and autocovariances, $R(u, \theta_0)$, $\theta_0 \in \Theta$ and Θ is compact. If $\theta \neq \theta'$ the set $\{\lambda | f_y(\lambda, \theta) = f_y(\lambda, \theta')\}$ is supposed to have positive Lebesgue measure.*

Let $\hat{\theta}$ be the estimated parameters which maximise log-likelihood function given in equation 2.12. Then $\hat{\theta}$ is an efficient estimator of θ_0 and, as $n \rightarrow \infty$,

$$\sqrt{n} (\hat{\theta} - \theta_0) \rightarrow_d N(0, \Gamma^{-1}(\theta_0))$$

where $\Gamma(\theta) = (\Gamma_{ij}(\theta))$ with

$$\Gamma_{ij}(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{\partial \log(f_{\theta}(\lambda))}{\partial \theta_i} \right) \left(\frac{\partial \log(f_{\theta}(\lambda))}{\partial \theta_j} \right) d\lambda.$$

This theorem shows the ML estimates are \sqrt{n} consistent, efficient and asymptotically Normally distributed. However, the log-likelihood function given in equation 2.12 requires the calculation of both the determinant and the inverse of Γ_{θ} and can thus be slow to compute. Use of the Durbin-Levinson algorithm, described in section 1.5, can speed up the computational time by expressing the log-likelihood in terms of the $\{\nu_m\}$ as

$$L(\theta) = -\frac{1}{2} \left(\sum_{m=1}^n \log \nu_{m-1} + \sum_{t=1}^n \frac{\varepsilon_t^2}{\nu_{t-1}} \right),$$

where $\varepsilon_t = y_t - \hat{y}_t$ and \hat{y}_t is the best linear predictor of y_t given $\{y_s\} 1 \leq s < t$, see section 1.4. The numerical complexity of this for a FARIMA process is $O(n^2)$, see, for example, Ammar [1998].

Since Γ_{θ} is a symmetric positive definite matrix, the Cholesky decomposition method can also be used to speed up computations by writing

$$\Gamma_{\theta} = U'U,$$

where U is an upper triangular matrix. This gives $\det \Gamma_{\theta} = (\det U)^2 = \prod_{j=1}^n u_{jj}^2$, where u_{jj} is the element of U in the j th row and j th column. The inverse of U is also simpler to find than that of Γ_{θ} . However, the numerical complexity of this method is still $O(n^3)$, see Press et al [1992].

The Whittle approximation to the log-likelihood function has numerical complexity of $O(n \log_2(n))$ and can thus greatly reduce computational time. The Whittle method is to express Γ_{θ} approximately in terms of the spectral density. The following approximations are made

$$\begin{aligned} \frac{1}{2n} \log \det \Gamma_{\theta} &\approx \frac{1}{4\pi} \int_{-\pi}^{\pi} \log(2\pi f(\lambda, \theta)) d\lambda \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \log(f(\lambda, \theta)) d\lambda + A, \end{aligned}$$

where $A = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log(2\pi) d\lambda$ is a constant, and

$$\frac{1}{2n} \mathbf{y}' \Gamma_{\theta}^{-1} \mathbf{y} \approx \sum_{k=1}^n \sum_{j=1}^n y_k y_j \left\{ \frac{1}{8\pi^2 n} \int_{-\pi}^{\pi} f^{-1}(\lambda, \theta) \exp(i\lambda(k-j)) d\lambda \right\}$$

$$= \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda, \theta)} d\lambda$$

where $I(\lambda)$ is the periodogram defined in section 1.7. The Whittle likelihood approximation is thus

$$L^*(\theta) = -\frac{1}{4\pi} \left(\int_{-\pi}^{\pi} \log(f(\lambda, \theta)) d\lambda + \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda, \theta)} d\lambda \right),$$

where the constant A does not effect the maximisation and is thus ignored. Further approximation of the integrals by Riemann sums gives the Whittle likelihood approximation in a discrete form

$$L_w(\theta) = -\frac{1}{2n} \left(\sum_{j=1}^n \log(f(\lambda_j, \theta)) + \sum_{j=1}^n \frac{I(\lambda_j)}{f(\lambda_j, \theta)} \right). \quad (2.13)$$

The use of this approximation instead of the exact log-likelihood function has been found by Fox and Taqqu [1986] and Dahlhaus [1989] to give estimates of the parameters with the same asymptotic distribution as the ML estimates. Further work by Giraitis and Surgailis [1990] proves the Whittle likelihood estimates still have this asymptotic distribution when the assumption of Gaussianity is relaxed and replaced with the assumption that the innovations are independent and identically distributed with finite fourth cumulant. Hence, when the order of the FARIMA(p,d,q) model is correctly specified the Whittle likelihood gives estimates which are \sqrt{n} consistent, efficient and asymptotically Normally distributed and these can be computed much faster than the ML estimates. The question of order selection is now considered.

Beran et al. [1998] consider FAR(p, d) models with p finite and $d \in (-0.5, \infty)$, that is FARIMA(p,d,0) models. They derive a suitable version of the Akaike information criterion, AIC, for this class of processes and show that it is of the same form as in the standard short memory situation, but with d treated as an additional parameter.

$$\text{AIC}(p) = -2 \log L(\tilde{A}) + 2(p + 1)$$

Studying the asymptotic sampling properties of the order selected by AIC, Beran et al [1998] show that, as in the short memory case, AIC does not provide a consistent order selection procedure for this class of processes, with the probability of underestimated p tending to zero as $n \rightarrow \infty$, but the probability of overestimating p tending to a non-zero finite constant. They then continue to show the corresponding versions of the BIC and IIC criteria, see Schwarz [1978] and Hannan and Quinn [1979], are consistent for this class of processes, where

$$\text{BIC}(p) = -2 \log L(\tilde{A}) + \log(n)(p + 1)$$

and, for $c > 1$,

$$\text{HIC}(p) = -2 \log L(\tilde{A}) + (2c \log \log n)(p + 1).$$

The extension of these results to include FARIMA(p,d,q) models, with $q \neq 0$, has not currently been theoretically proven. However, extensive simulation studies have been carried out by, for example, Schmidt and Tcherning [1995], Crato and Ray [1996] and Bisaglia [2002] which indeed seem to show that use of AIC tends to lead to inconsistent estimates of the order, with the model chosen often being overparametrised, whilst the use of BIC and HIC select the correct model with a higher frequency.

If the fitted model is misspecified, or indeed if a 'true' FARIMA(p,d,q) model does not exist, the estimates of d and the short memory parameters using either the ML or the Whittle likelihood methods can have large biases and variances, see, for example, Taquq and Teverovsky [1996], Bisaglia and Bordignon [2002] and Smith et al. [1997]. This has led to the development of *semi-parametric* and *non-parametric* methods of estimating d which are more robust to model misspecification. Some of these methods are discussed below.

2.3.2 Geweke and Porter-Hudak (GPH) method

The GPH method was originally proposed by Geweke and Porter-Hudak [1983]. It is motivated by the observation that although classical ARMA and ARIMA models can capture the behaviour of a FARIMA process at higher frequencies, where the spectral density is bounded above and below, they fail to do so at low frequencies. They thus suggest looking first at these low frequencies to estimate d , then to filter out this d and fit an ARMA model to the remaining series $\{\hat{x}_t\}$. As such, the estimate of d does not rely on a specific model being fitted and is thus considered a *non-parametric* method of estimation.

As mentioned in section 2.2, when long memory of a series, $\{y_t\}$, is the result of fractionally differencing a short memory process, $\{x_t\}$, the relationship between the spectral densities of $\{y_t\}$ and $\{x_t\}$ can be defined as

$$f_y(\lambda) = (4 \sin^2(\frac{\lambda}{2}))^{-d} f_x(\lambda),$$

where $f_x(\lambda)$ is a bounded continuous function which tends to a constant as $\lambda \rightarrow 0$. This form of spectral density is common to FARIMA and FExp processes and also general fractional Gaussian noise, see section 2.2.

Geweke and Porter-Hudak [1983] suggest taking the logs of this to give

$$\log(f_y(\lambda)) = \log(f_x(0)) - d \log(4 \sin^2(\frac{\lambda}{2})) + \log(\frac{f_x(\lambda)}{f_x(0)}). \quad (2.14)$$

Adding $\log(I(\lambda_j))$ to both sides of equation 2.14 and rearranging gives

$$\log(I(\lambda_j)) = \log(f_x(0)) - d \log(4 \sin^2(\frac{\lambda_j}{2})) + \log(\frac{f_x(\lambda_j)}{f_x(0)}) + \log(\frac{I(\lambda_j)}{f_y(\lambda_j)}), \quad (2.15)$$

Considering only frequencies near $\lambda = 0$, this gives $f_x(\lambda) \approx f_x(0)$ and therefore $\log(\frac{f_x(\lambda_j)}{f_x(0)})$ can be considered negligible. The similarity between equation 2.15 and the standard linear regression equation is then pointed out and the suggested GPH estimate for d is obtained via a standard least squares approach,

$$\hat{d} = \frac{\sum_{j=1}^m (-\log(4 \sin^2(\frac{\lambda_j}{2}) - \hat{\mu}) \log(I(\lambda_j)))}{\sum_{j=1}^m (-\log(4 \sin^2(\frac{\lambda_j}{2}) - \hat{\mu})^2)}, \quad \hat{\mu} = \frac{1}{m} \sum_{j=1}^m -\log(4 \sin^2(\frac{\lambda_j}{2})), \quad (2.16)$$

where m is the number of frequencies to include. Geweke and Porter-Hudak [1983] prove the consistency of this estimate for $d < 0$. Robinson [1995a], studies the asymptotic behaviour of the periodogram of long-memory processes and gives the asymptotic distribution of a related estimate $d(l)$, in which the initial l frequencies are not included, thus only frequencies λ_j for $j \in (l, \dots, m)$ are included.

Hurvich, Deo and Brodsky [1998] further study the asymptotic distribution of the GPH estimate of d . They relax the requirement of an increasing lower truncation number l and give results for the original GPH estimate defined in equation 2.16 with the summations over $j \in (1, \dots, m)$. The following assumptions are made by Hurvich, Deo and Brodsky [1998].

Assumption 2.3.1. The series, y_t , $t \in (1, 2, \dots)$ is a Gaussian process, with spectral density $f_y(\lambda)$.

Assumption 2.3.2. Let

$$f_y(\lambda) = (4 \sin^2(\frac{\lambda}{2}))^{-d} f_x(\lambda),$$

such that

$$f'_x(0) = 0, f''_x(\lambda) < C < \infty, \text{ and } f'''_x(\lambda) < C < \infty$$

for some positive constant C and all λ in a neighbourhood of zero.

Assumption 2.3.3. $m \rightarrow \infty, n \rightarrow \infty$, with $m/n \rightarrow 0$ and $(m \log m)/n \rightarrow 0$.

These assumptions are similar to those of Robinson [1995a], however the conditions on l have been removed. Theorem 2.3.2 below uses these to give the asymptotic bias, variance and MSE of the GPH estimate \hat{d} . Theorem 2.3.3 then gives the asymptotic distribution of \hat{d} under slightly stronger conditions on m . These correspond to Theorems 1 and 2 of Hurvich, Deo and Brodsky [1998].

Theorem 2.3.2. *Let y_t be a series of length n . Let assumptions 2.3.1 - 2.3.3 hold, with $d = d_0$. Let \hat{d} be the least squares estimates of d_0 , obtained from equation 2.16, with the summations taken over $j \in (1, \dots, m)$. Then, as $n \rightarrow \infty$,*

$$E(\hat{d} - d_0) = \frac{-2\pi^2 f_x''(0) m^2}{9 f_x(0) n^2} + o\left(\frac{m^2}{n^2}\right) + O\left(\frac{\log(m)^3}{m}\right),$$

$$\text{Var}(\hat{d}) = \frac{\pi^2}{24m} + o\left(\frac{1}{m}\right)$$

and

$$\begin{aligned} \text{MSE}(\hat{d}) = E\left((\hat{d} - d_0)^2\right) &= \frac{4\pi^4}{81} \left(\frac{f_x''(0)}{f_x(0)}\right)^2 \frac{m^4}{n^4} + \frac{\pi^2}{24m} + o\left(\frac{1}{m}\right) \\ &+ o\left(\frac{m^4}{n^4}\right) + O\left(\frac{m \log(m)^3}{n^2}\right), \end{aligned}$$

Theorem 2.3.3. *Let y_t be a series of length n . Let assumptions 2.3.1 - 2.3.3 hold, with $d = d_0$. Let \hat{d} be the least squares estimates of d_0 , obtained from equation 2.16, with the summations taken over $j \in (1, \dots, m)$. As $n \rightarrow \infty$, let $m = o(n^{4/5})$ and $\log(n)^2 = o(m)$, then*

$$\sqrt{m}(\hat{d} - d_0) \rightarrow_d N\left(0, \frac{\pi^2}{24}\right)$$

Theorem 2.3.3 shows the GPH estimate of d is \sqrt{m} consistent. Since $m = o(n^{4/5})$, this estimate is less efficient than the parametric ML estimate of Whittle likelihood estimates given in section 2.3.1, which are \sqrt{n} consistent. However, the GPH estimate is robust to changes in the short memory component. From Theorem 2.3.2 it can be seen that increasing m increases the bias of the estimate whilst decreasing the variance, thus the choice of m produces a trade off. Geweke and Porter-Hudack [1983] suggested m be taken as $n^{0.5}$, based on a simulation study. Empirical findings of Crato and Lima [1994] and Porter-Hudak [1990] suggest that an m be taken in the range $n^{0.5}$ to $n^{0.7}$. Hurvich, Deo and Brodsky [1998] suggest an

optimal choice of m as the value of m which minimises the MSE given in Theorem 2.3.2, that is

$$m^* = \left(\frac{27}{128\pi^2} \right)^{1/5} \left(\frac{f_x(0)}{f_x''(0)} \right)^{2/5} n^{4/5}.$$

However, this choice of m requires further knowledge of $f_x(\lambda)$ and is $O(n^{4/5})$ whilst the asymptotic normality of the estimate requires $m = o(n^{4/5})$, see Theorem 2.3.3. Hurvich and Beltra [1994] consider a data driven choice of m , based on a method of cross validation. They define $\hat{d}^{(-j)}$ as the GPII estimate of d , leaving out the frequency at j . They then define

$$\hat{f}_j^{(-j)} = \exp \left\{ \frac{1}{p-1} \sum_{k=1, k \neq j}^p \left[\log \left(\frac{I_k}{|1 - \exp(-i\lambda_k)|^{-2\hat{d}^{(-j)}}} \right) + C \right] \right\} \\ \times |1 - \exp(-i\lambda_j)|^{-2\hat{d}^{(-j)}},$$

where $C = 0.577216\dots$ is Euler's constant. They then take the choice of m to be that which minimises either the frequency domain cross validation, see Wahba and Wold [1975] and Hurvich [1985],

$$FDCV_p(m) = \frac{1}{p} \sum_{j=1}^p \left(\log(\hat{f}_j^{(-j)}) - \log(I_j) - C \right)^2 - \pi^2/6$$

or cross-validated log likelihood, see Beltrao and Bloomfield [1987],

$$CVLL_p(m) = \frac{1}{p} \sum_{j=1}^p \left(\log(\hat{f}_j^{(-j)}) + \frac{I_j}{\hat{f}_j^{(-j)}} \right).$$

Although the simulation studies of Hurvich and Beltra [1994] show the CVLL method of choosing m generally gives estimates which outperform those based on arbitrarily chosen values of m , the theoretical proofs of the behaviour of m or \hat{d} using these are not given. It should also be noted that use of these data driven criteria require another user chosen parameter p , although Hurvich and Beltra [1994] say the chosen m is not very sensitive to p .

2.3.3 The Local Whittle Method

The local Whittle method of estimation for the long-memory parameter d was suggested by Künsch [1987], although he did not establish any statistical properties. Robinson [1995b] later studied the estimate and derived its asymptotic distribution.

Similar to the non-parametric GPII method it uses only the lower frequencies of the spectral density to estimate d , requiring that

$$f(\lambda) \sim G\lambda^{-2d} \quad \text{as } \lambda \rightarrow 0+,$$

for $G \in (0, \infty)$ and $d \in (-0.5, 0.5)$. As with the GPII method, FARIMA, FExp and general fGN processes all satisfy this assumption, thus the method is more robust to changes in the model than the parametric methods. The suggestion is to then use this form of the spectral density in the Whittle likelihood approximation, see equation 2.13, for frequencies $j \in (1, \dots, m)$. This gives the local Whittle estimate of d as the estimate of d which minimises the objective function $Q(d, G)$,

$$Q(d, G) = \frac{1}{m} \sum_{j=1}^m \left\{ \log G\lambda_j^{-2d} + \frac{\lambda_j^{2d}}{G} I_j \right\}$$

where $I(\lambda)$ is the sample spectrum defined in equation 1.45. Since the local Whittle method uses assumptions similar to the GPII method it will also be regarded as a non-parametric method of estimating d . Now, an estimate of G can be found in terms d as

$$\hat{G}(d) = \frac{1}{m} \sum_{j=1}^m \lambda_j^{2d} I_j,$$

and substitution of this into $Q(d, G)$ gives

$$\hat{d} = \arg \min R(d)$$

where

$$R(d) = \log \hat{G}(d) - (2d) \frac{1}{m} \sum_{j=1}^m \log(\lambda_j). \quad (2.17)$$

Robinson [1995b] proves the asymptotic normality of the local Whittle estimate under several assumptions. Assumption A1' of Robinson [1995b] puts a rate of convergence of the spectral density to the required form. Assumption A2' connects this rate of convergence to the maximum choice of m . The weakest condition on m is given when $\beta = 2$, which gives $m = o(n^{4/5})$. Note, this is equivalent to the condition placed on m to prove the asymptotic Normality of the GPII estimate. Assumption A3' is weaker than the Gaussianity assumption required for the GPII method. Under these assumptions, Theorem 2 of Robinson [1995b] proves the following result.

Theorem 2.3.4. *Let \hat{d} be the local Whittle estimate of d_0 found by minimising the objective function given in equation 2.17. Let the assumptions of Robinson [1995b] hold. Then, as $n \rightarrow \infty$,*

$$\sqrt{m}(\hat{d} - d_0) \sim N\left(0, \frac{1}{4}\right).$$

Similar to the GPII estimate, Theorem 2.3.4 shows the local Whittle estimate is \sqrt{m} consistent for d . However, comparison of Theorem 2.3.4 with Theorem 2.3.3 shows, for the same choice in m , the asymptotic variance of the local Whittle estimate to be approximately 0.6 of the asymptotic variance of the GPH method.

As with the GPII method, the choice of m is a trade off between bias and variance. A study carried out by Bhansali and Kokoszka [2001a] suggested a suitable range of m to be $(n/100, n/40)$. A study of the effect of using different band lengths was also carried out by Taqqu and Teverovsky [1996] where the trade-off between variance and bias was found to be optimal between a similar range of frequencies. The FDCV and CVLL methods discussed in section 2.3.2 could also be used.

2.3.4 The FExp method

The fractional exponential, FExp, set of models were introduced in section 2.2.3. A parametric approach to finding the estimate of d could be applied to these models, using either the exact likelihood or Whittle likelihood approximation, see section 2.3.1. The method of estimating d discussed in this section is a semiparametric method. It was first discussed by Janacek [1982], see also Robinson [1994a], though they provided no theoretical work. It was later studied by Moulines and Soulier [1999], [1998] and Hurvich and Brodsky [2001].

As opposed to the local Whittle and GPII methods, which use a decreasing amount of the spectral density at low frequencies, the FExp method finds an estimate of d from use of the full spectrum. Moulines and Soulier [1999] thus refer to it as a *broadband* estimate. Since all of the frequencies are to be used in estimation, the difficult question of how many frequencies to include that arose for the GPII and local Whittle methods is no longer applicable.

For an observed time series, $\{y_t\}$, the spectral density is assumed to be of the form

$$\log(f(\lambda)) = -d \log(2 \sin^2 \frac{\lambda}{2}) + \sum_{j=0}^{\infty} c_j \cos(j\lambda),$$

with $d \in (-0.5, 0.5)$ and coefficients $c_j \in \mathbb{R}$ for $j \in (0, 1, \dots)$. In general, this expansion has infinitely many nonzero coefficients. A nonparametric estimate of the spectral density is found by truncating this infinite summation at some finite s such

that the finite series provides a reasonable approximation to the true spectral density. Once a suitable s has been chosen, the coefficients along with d can be estimated. Moulines and Soulier [1999] make use of linear regression similar to the GPI method to carry out the estimations, though suggest also that a Whittle likelihood approach could be applied.

To reduce variance, the periodogram, $I(\lambda_j)$, is first smoothed in a method suggested by Robinson [1995a]. The periodogram is first divided into nonoverlapping segments of equal size m and then averaged over each of these segments to produce Y_k ,

$$Y_k = \log \left(\exp(-\psi(m)) \sum_{i \in J_k} I(\lambda_i) \right),$$

where $k = 1, \dots, K$ is the segment number, $J_k = (m(k-1) + 1, \dots, mk)$ and $\psi(m) = \Gamma'(m)/\Gamma(m)$ is the digamma function added to remove bias. The FExp estimator, \hat{d} , is then the least squares estimator

$$\hat{d} = \arg \min_{\bar{d}, \bar{c}_0, \dots, \bar{c}_s} \sum_{k=1}^K \left(Y_k - \bar{d}g(\lambda_k) - \sum_{j=0}^s \bar{c}_j h_j(\lambda_k) \right)^2 \quad (2.18)$$

where $\lambda_k = (2k-1)\pi/2K$, $h_j(x) = \cos(jx)$ and $g(x) = \log(2 \sin^2 \frac{x}{2})$. Moulines and Soulier [1999] give the following Theorem of this estimate under several conditions, including that the process is Gaussian.

Theorem 2.3.5. *Let $\{y_t\}$ be an observed time series of length n , let \hat{d} be the FExp estimate of d given in equation 2.18 and let the assumptions of Moulines and Soulier [1999] hold. Then, as $n \rightarrow \infty$,*

$$\sqrt{\frac{n}{s}} (\hat{d} - d) \rightarrow_d N(0, m\psi'(m))$$

and

$$\left(\frac{n}{s}\right) E \left((\hat{d} - d)^2 \right) \rightarrow m\psi'(m).$$

Moulines and Soulier [1999] show, under the conditions of the proof, the best rate of convergence of the FExp estimate to the true parameter is that the MSE is $O(n^{-2\beta/(1+2\beta)})$ for a constant $\beta > 1/4$ which is dependent on the rate of decay of the short memory coefficients $\{c_j\}$. For $\beta > 1/4$ the estimate is at least $\sqrt{n^{4/5}}$ consistent. As β becomes arbitrarily large, the estimate approaches the \sqrt{n} consistency of the parametric methods. In particular, if the coefficients $\{c_j\}$ decay at an exponential

rate, Moulines and Soulier [1999] show s can be taken as order $\log(n)$ and the FExp estimate is thus $\sqrt{n/\log(n)}$ consistent.

The FExp method of estimating d depends on the choice of s and m . The value for m relates to the amount of smoothing of the periodogram and presents a trade-off between bias and variance. Moulines and Soulier [1999] set their choice of m to 4, though they suggest that a different choice of m will have little effect.

The choice of s can be found using a suitable criterion. The use of Mallows's C criterion was suggested by Robinson [1994a], without theoretical justification. This gives the choice of s decided by

$$\hat{s} = \arg \min_{1 \leq s \leq K} \frac{\pi}{K} \sum_{k=1}^K \left(Y_k - \bar{d}g(\lambda_k) - \sum_{j=0}^s \bar{c}_j h_j(\lambda_k) \right)^2 + \frac{4\pi s \sigma_m^2}{K} \quad (2.19)$$

where $\sigma_m^2 = \text{var}(x_{2m}^2) = \psi'(m)$, $\lambda_k = (2k-1)\pi/2K$, $h_j(x) = \cos(jx)$ and $g(x) = \log(2 \sin^2 \frac{x}{2})$.

Moulines and Soulier [1998] prove that \hat{s} is asymptotically optimal in terms of minimising the MSE. Theorem 3, and remarks thereafter, of Moulines and Soulier [1998] give the following theorem.

Theorem 2.3.6. *Let $\{y_t\}$ be an observed time series of length n , let \hat{s} be the choice of s found by minimising Mallows's C criterion, see equation 2.19, let \hat{d}_s be the FExp estimate of d given in equation 2.18 using $s = \hat{s}$ and let the assumptions of Moulines and Soulier [1999] hold. Then, as $n \rightarrow \infty$,*

$$\hat{s} = O(n^{1/(1+2\beta)})$$

and

$$E \left(\left(\hat{d}_s - d \right)^2 \right) = O(n^{-2\beta/(1+2\beta)}).$$

Theorem 2.3.6 shows the estimate \hat{d}_s is $\sqrt{n^{2\beta/(1+2\beta)}}$ consistent, which, as previously discussed, is the best rate of convergence for the FExp estimate. The choice of \hat{s} is thus asymptotically optimal.

2.3.5 FAR Method

The FAR method of estimating d was introduced and studied by Bhansali and Kokoszka [2001a] and Bhansali et al [2006]. It is a semiparametric method similar in approach to the semiparametric FExp method. The observed series, $\{y_t\}$, is assumed to be taken from a process of the form

$$(1 - L)^d y_t = x_t \quad (2.20)$$

where

$$x_t = e_t + \sum_{j=1}^{\infty} \phi_j x_{t-j}, \quad (2.21)$$

with $d \in (-0.5, 0.5)$. Brillinger [1981] shows such a form can be found for $\{x_t\}$ provided it has a nonzero spectral density, $f_x(\lambda)$, and ACF which decay such that

$$\sum_{u=-\infty}^{\infty} (1 + |u|^\beta) |\tau(u)| < \infty.$$

The FAR method is thus applicable to a wide range of long memory processes, including FARIMA, FExp and general fGN. As with the FExp approach, for a given time series of length n , the summation in equation 2.21 is truncated at a finite p . The parameters d and $\{\phi_j\}$ $j \in (1, \dots, p)$ are then estimated by maximising the Whittle likelihood function, see section 2.3.1.

Under several assumptions, Bhansali et al [2006] prove the asymptotic normality of the FAR estimate of d . Unlike the results for the FExp method, this Theorem is valid for non-Gaussian processes.

Theorem 2.3.7. *Let $\{y_t\}$ be an observed time series of length n , let \hat{d} be the FAR estimate of d and $\hat{\phi}_j$ be the FAR estimates of ϕ_j for $j \in (1, \dots, J)$, $J = o(p)$. Let the assumptions of Bhansali et al [2006] hold. Then, as $n \rightarrow \infty$,*

$$\left(\sqrt{\frac{n}{p}} \right) (\hat{d} - d) \rightarrow_d N(0, 1)$$

and, for $j \in (1, \dots, J)$,

$$\left(\sqrt{\frac{n}{p}} \right) (\hat{\phi}_j - \phi_j) \rightarrow_d \tau_j N(0, 1)$$

where

$$\tau_j = \sum_{k=1}^j \frac{\phi_{(j-k)}}{k}.$$

To obtain parameter free limits, Bhansali et al [2006] also show that an estimate of τ_j can be found by

$$\hat{\tau}_j = \sum_{k=1}^j \frac{\hat{\phi}_{(j-k)}}{k},$$

and

$$\hat{\tau}_j \rightarrow \tau_j, \quad \text{as } n \rightarrow \infty.$$

Comparison of the asymptotic distribution of the FAR estimate shown in Theorem 2.3.7 with the asymptotic distribution of the FExp estimate shown in Theorem 2.3.5 shows if p and s are chosen equally, the rate of convergence for both will be the same. For example, if $p = s = \log(n)$, then both estimates will be $\sqrt{n/\log(n)}$ consistent. The variance of the FExp estimate, in this case, will be asymptotically $m\psi'(m)$ times that of the FAR estimate, for some user chosen m , and since $m\psi'(m) \in (1, \pi^6/6]$ for $m \in \mathbb{N}$, the FAR estimate is the more efficient.

Bhansali and Kokoszka [2001a] also prove the consistency of the FAR estimates for a class of α -stable processes. The following result corresponds to Theorem 4.1 of Bhansali and Kokoszka [2001a]

Theorem 2.3.8. *Let $\{y_n\}$ be an observed time series of length n , let $\hat{\Theta}$ be the FAR estimate of Θ . Let the assumptions of Bhansali and Kokoszka [2001a] hold. Then, for $d \in (0, 1 - 1/\alpha)$, as $n \rightarrow \infty$,*

$$\hat{\Theta} \rightarrow_p \Theta.$$

Theorem 2.3.8 widens the range of processes for which the FAR method of estimating d is applicable to include α -stable processes which have innovations with infinite variance.

Application of the FAR method requires a value of p to be fitted for the given time series. Simulation studies by Bhansali and Kokoszka [2001a] suggest use of criterion such as AIC or BIC. The results using both criterion appear to give consistent estimates of d for a range of FARIMA(p, d, q) processes with $p, q \in (0, 1, 2)$ and both Gaussian and α -stable innovations, however the estimates found using BIC tend to dominate those of AIC.

theoretical justification of the use of AIC or BIC, equivalent to Theorem 2.3.6 for using Mallows's C criterion for the FExp method, have not yet been established.

3. CHAOS

This chapter introduces the concept of *chaos*, with particular focus on a range of chaotic intermittency maps which can be used to produce non-linear non-Gaussian time series with long memory properties. A *chaotic system*, as defined, for example, by Hilborn [2000], is one which, although purely deterministic, is aperiodic (never exactly repeats) and can appear to have random behaviour.

Evidence of chaos has been found in a wide variety of sources. Moon and Holmes [1979] found chaotic behaviour in the strain of a steel beam oscillating between two magnets. Shaw [1984] found the behaviour of light through dripping water to be chaotic. Applications of chaos in biology can be found in Glass and Mackey [1988]. Lawrance and Balakrishna [2001] discuss how chaotic maps can be designed to have autocorrelations suitable for application to communications. Further applications in areas such as acoustics, chemistry, circuits, lasers and plasmas can be found in Lauterborn [1981], Simoyi et al [1982], Rollins and Hunt [1984], Mork et al [1990] and Sagdeev et al [1990], respectively.

One of the first papers concerning a chaotic system was Hadamard [1898] which shows that billiard trajectories are unstable and diverge from one another. Around the same time, Poincaré [1895] discovered differential equations with chaotic solutions and his work laid the foundations for chaos theory. However, it was not until the arrival of computers several decades later that chaos really began to be popular and as such it is a relatively new area. One definition of chaos, introduced by Devaney [1986], states the three conditions for chaos as

1. Sensitivity dependence.
2. Transitivity.
3. A dense set of periodic points.

These conditions will be discussed in greater detail in section 3.1.

The class of chaotic intermittency maps produce time series which, in addition to showing long memory properties, seem to switch between periods of smooth behaviour and periods of chaotic behaviour. As such, they have been considered for modelling internet traffic, see Mondragon [1999] and Bhansali et al [2005], which switches between ON/OFF states. Other work concerning the binary discretising of chaotic maps for communications can be seen in Lawrance and Wolff [2003]. Bhansali et al [2006] also considers use of these chaotic intermittency maps to model financial returns, which have also been found to exhibit long memory and switch between

periods of relative inactivity and periods of high volatility.

This range of chaotic intermittency maps are defined in section 3.2. An investigative simulation study on the rate of decay of the correlations of the series produced by these chaotic intermittency maps will then be carried out in chapter 4. The results from these studies motivates the new dual parameter long memory model introduced in Chapter 5, where an application of these new methods to the chaotic intermittency maps is also carried out. New stochastic extensions of these maps are discussed in Chapter 6.

3.1 Basic Properties

The aim of this section is to introduce some fundamental properties of chaotic maps relevant to the study of chaotic intermittency maps which will be introduced in section 3.2. As such, the focus is on properties of discrete chaotic series generated from chaotic maps. Details specific to continuous or multivariate chaotic series are not discussed and several other topics are also omitted, such as strange attractors, bifurcation diagrams, Feigenbaum numbers and crises, as they will not be used in the study of the chaotic intermittency maps. For further details of these and other topics regarding chaos see reviews such as Hilborn [2000], Ott [2002], Szemplinska-Stupnicka [2003] and Smith [2007].

Now, let $\{w_t, t \in \mathbb{N}\}$ be a deterministic sequence, depending only on the initial condition w_0 , produced by iteratively applying a one-dimensional map, ζ , such that

$$w_{t+1} = \zeta(w_t).$$

The sequence $\{w_t\}$ is known as an *orbit* of the map ζ . If the map is chaotic, taking the definition of Devaney [1986], this orbit will possess the properties presented in the following subsections. Note, the sequence need only be deterministic in one direction, such that given only w_t , w_{t+1} can be determined, although many values for w_{t-1} may be possible, see Lawrance [1991], [1992] and Lawrance and Spencer [1998]. For the chaotic intermittency maps generally two possible values of w_{t-1} are present for w_t , but these stochastic reverse processes are not considered here.

3.1.1 Sensitivity Dependence

A defining attribute of a chaotic map is that its orbits possess exponentially sensitive dependence on initial conditions.

Consider two initial conditions w_0 and $v_0 = w_0 + \delta_0$, such that the resulting orbits are $\{w_t\} \in J$ and $\{v_t = w_t + \delta_t\} \in J$ with $0 < \delta_t$ for $t \in \mathbb{N}$ and J is a closed interval of real numbers.

Also let the condition that $w_t \neq v_j$ for all $t, j \in \mathbb{N}$ hold. Note, if $w_t = v_j$ for some $t, j \in \mathbb{N}$ it implies $w_{t+k} = v_{j+k}$ for all $k \in \mathbb{N}$ since the maps are deterministic.

Now, if the orbits possess exponentially sensitive dependence on initial conditions then for large t

$$\lim_{\delta_0 \rightarrow 0} \frac{\delta_t}{\delta_0} = e^{\lambda t} \quad (3.1)$$

for some $\lambda > 0$. This implies that orbits from a chaotic map with initial conditions arbitrarily close will diverge at an exponential rate. The condition that the orbits remain within a bounded interval J removes the trivial cases of unbounded maps, such as $w_t = w_0^t$ with $w_0 > 1$.

The variable λ is known as *the Lyapunov Exponent*. A positive value for the Lyapunov Exponent, as shown in equation 3.1, leads to exponentially sensitive dependence on initial conditions. Alternatively, a negative value for the Lyapunov Exponent would imply that neighbouring orbits converge, whilst a Lyapunov Exponent equal to zero implies the order of the error terms stays constant.

In finite time the amount of sensitivity to initial conditions may depend on the initial conditions. Some orbits may diverge at different rates. As such, it is important to distinguish between the *global Lyapunov Exponents*, which are constant and show the general properties of the map, showing the rate of divergence that eventually most typical orbits will possess, and the *finite time Lyapunov Exponents*, which vary according to the local behaviour of the maps.

The sensitivity of initial conditions for chaotic maps means that, although the orbits are completely deterministic and bounded, meaningful long term prediction is essentially impossible.

Another important point is that the sensitivity to initial conditions calls into question the accuracy of any simulated orbits. The simulated orbit will diverge away from the 'true' orbit at an exponential rate and it may then seem questionable that the simulated orbit represents a 'true' orbit in any way. This is a far from trivial problem and leads to the important topic of *shadowing*.

Shadowing theory states that although the simulated orbit may diverge from the 'true' orbit at an exponential rate, there exists another 'true' orbit with slightly different initial conditions that remains close to the simulated orbit for a long period of time. Thus the simulated orbit has essentially the same properties as a 'true' orbit and simulation studies are meaningful.

The existence of shadowing orbits for invertible hyperbolic maps was established by Anosov [1967] and Bowen [1970]. Early discussion of the existence of shadowing orbits for noninvertible nonhyperbolic maps can be found in Grebogi et al [1990] which gives the basic 'rule of thumb' that given an initial error δ_0 , and labeling the distance between the shadowing orbit and the simulated orbit at time t as η_t , then $\eta_t \leq \sqrt{\delta_0}$ for $t \in (1, \dots, N)$, where $N \sim 1/\sqrt{\delta_0}$. The point at time N where the shadowing orbit finally diverges from the simulated orbit is known as a *glitch*. Work by Dawson et al [1994] showed how this 'rule of thumb' can break down and that the existence of shadowing orbits for any meaningful length of time could not be

found for some maps with finite time Lyapunov Exponents which fluctuated around zero. This led to the idea that the existence of the shadowing orbits was dependent on the finite time Lyapunov Exponents and Sauer et al [1997] established that the distribution, $p(y)$, of the log shadowing distances, $y_t = \log(\eta_t)$, is exponential with

$$p(y) = \frac{2|m|}{\sigma^2} e^{-2|m|(y - \log(\delta_0))/\sigma^2}$$

and the expected time between glitches, N , is

$$E(N) \sim \delta_0^{-2|m|/\sigma^2} \quad (3.2)$$

where m is the mean of the finite time Lyapunov Exponent and σ is its standard deviation. It can be seen from equation 3.2 that as $|m|/\sigma^2 \rightarrow 0$, the expected time between glitches can be small regardless of the size of δ_0 , thus in this case a shadowing orbit may not be found and results from simulations may not represent the true properties of the maps.

Even when the existence of a shadowing orbit has been established it has been remarked that this shadowing orbit may not be typical of the map, see for example Quinlan and Tremaine [1992]. If the shadowing orbit is atypical, the simulated orbit may not possess the typical properties of 'true' orbits from the map. The important question of under what conditions shadowing orbits are typical or atypical remains open. A detailed survey of shadowing methods can be found in Hayes and Jackson [2005].

In summary, although deterministic, the sensitive dependence of initial conditions displayed by orbits from chaotic maps leads to an element of uncertainty which is producing increasing interest in the topic of chaos for statisticians.

3.1.2 Transitivity

According to Devaney [1986], the second defining property of a chaotic map is *transitivity*.

Define a map $\zeta: J \rightarrow J$. The map, ζ , is said to be *transitive* if for every pair of subintervals $A, B \subseteq J$ there exists an n such that

$$\zeta^n(A) \cap B \neq \emptyset,$$

where ζ^n represents n iterations of the map ζ .

Allowing the size of the subinterval A to tend to zero, this is related to the property of an orbit being *dense*. An orbit, $\{w_t\}$ is said to be *dense* if for every subinterval $B \subseteq J$ there is an element of $\{w_t\}$ in B . If a map has an orbit which is dense then it is transitive.

The transitivity of chaotic maps is related to the recurrence properties. Let $w_0 \in B \subseteq J$. Let $T_B(w_0)$ be the *recurrence time*, that is the time it takes for the orbit to first return to B . Now, define $H_B(n)$ as

$$H_B(n) = P(T_B(w_0) \leq n | w_0 \in B) \quad n \in (1, 2, \dots). \quad (3.3)$$

Due to the transitivity property of chaotic maps $H_B(n) \rightarrow 1$ as $n \rightarrow \infty$, that is given the initial condition w_0 is in the subinterval B , the orbit is expected to return to B eventually almost surely. Young [1999] shows the rate of convergence of $H_B(n)$ to 1 is related to the rate of decay of the autocorrelations. The following Theorem is a restatement of Theorems 2 and 3 of Young [1999].

Theorem 3.1.1. *Let $H_B(n)$ be defined by equation 3.3 for a chaotic map $\zeta : J \rightarrow J$. Then, as $n \rightarrow \infty$,*

a) *If $1 - H_B(n) = O(n^{-\alpha})$ for some $\alpha > 0$ and every $B \subset J$, then*

$$\text{Cov}(w_0, w_n) = O(n^{-\alpha})$$

b) *If $1 - H_B(n) = O(\theta^{-n})$ for some $\theta < 1$ and every $B \subset J$, then there exists $\tilde{\theta} < 1$ such that*

$$\text{Cov}(w_0, w_n) = O(\tilde{\theta}^{-n})$$

The results of this theorem were used in the proofs for finding the asymptotic rates of decay for the autocorrelations of the chaotic intermittency maps discussed in section 3.2 and hence showing that the orbits possess long memory behaviour.

3.1.3 Periodic Points

Another distinguishing feature of chaotic maps is the existence of a dense set of *periodic points*.

A *period- n point* of a map $\zeta : J \rightarrow J$ is a point w_0 such that

$$\zeta^n(w_0) = w_0, \quad (3.4)$$

where n is the smallest integer for which equation 3.4 holds. A *fixed point* is a point w_0 such that

$$\zeta(w_0) = w_0,$$

that is $n = 1$ in equation 3.4.

The behaviour of an orbit near a fixed point, with initial condition $w_0 + \delta_0$ and first iterate $w_0 + \delta_1$, say, can be seen by looking at the Taylor expansion of the map

$$\zeta(w_0 + \delta_0) = w_0 + \delta_1 = \zeta(w_0) + \zeta'(w_0)\delta_0 + o(\zeta'(w_0)\delta_0) \quad (3.5)$$

$$\Rightarrow \delta_1 = \zeta'(w_0)\delta_0 + o(\zeta'(w_0)\delta_0),$$

where $\zeta'(w_0)$ is the derivative of ζ evaluated at w_0 . From equation 3.5, it can be seen that if $|\zeta'(w_0)| < 1$ then $|\delta_1| < |\delta_0|$ and thus an orbit near the fixed point becomes closer to the fixed point after each iteration. This is known as a *stable fixed point*. The extreme case when $|\zeta'(w_0)| = 0$ is known as a *super stable fixed point*. Alternatively, if $|\zeta'(w_0)| > 1$ then $|\delta_1| > |\delta_0|$ and thus the orbit diverges away from the fixed point. This is known as an *unstable fixed point*. If $|\zeta'(w_0)| = 1$ then $|\delta_1| \approx |\delta_0|$. This is known as an *indifferent* or *neutral fixed point*.

For a neutral fixed point, an initial condition near the fixed point produces another point approximately the same distance from the fixed point. This produces a *laminar region* of the orbit, that is, a section of the orbit which remains smooth for possibly very long periods of time. The eventual repulsion or attraction of the orbit from this laminar region depends on the higher order terms of the Taylor expansion. Bhansali et al [2005] thus further classify neutral fixed points. If, for a neutral fixed point w_0 of a map ζ which is continuous near w_0 and small δ , $|\zeta'(w_0 \pm \delta)| > 1$, the orbit will eventually leave the laminar region and this is known as a *weakly repelling* neutral fixed point. If $|\zeta'(w_0 \pm \delta)| < 1$ the orbit will slowly converge to the fixed point and this is known as a *weakly attracting* neutral fixed point. The neutral fixed point, w_0 , is *attracting/repelling* if $|\zeta'(w_0 + \delta)| < 1$ while $|\zeta'(w_0 - \delta)| > 1$ for some small positive or negative δ . In this situation, the orbit's attraction or repulsion depends on which side of the fixed point it lies.

Note that the time it takes for an orbit to escape from a laminar region is related to recurrence times, in that they essentially measure the same thing, in that leaving the laminar region is equivalent to returning to non-laminar region.

If an orbit seems to switch back and forth between smooth behaviour and chaotic behaviour it is said to be *intermittent*, as is the case with orbits from the chaotic intermittency maps to be introduced in section 3.2. This form of intermittency was first studied by Manneville and Pomeau [1980]. Other forms of intermittency include orbits which seem to switch between periodic and chaotic behaviour and orbits which seem to switch between periodic and quasi-periodic behaviour.

In general, period- n points can be classified in a similar manner as fixed points. Take the map $\tilde{\zeta}_n = \zeta^n$, then, from the definition of period- n points given in equation 3.4, a period- n point of the map ζ is a fixed point of the map $\tilde{\zeta}_n$. *Stable, unstable* and *neutral period- n points* of the map ζ are thus those points which produce stable, unstable and neutral fixed points of the map $\tilde{\zeta}_n$ respectively.

Due to the required properties of sensitivity dependence and transitivity, discussed in sections 3.1.1 and 3.1.2, for a map to be chaotic, in terms of the definition by Devaney [1986], it must not possess any stable period- n points for any finite value of n . If a map is transitive, the probability of a typical orbit coming within a small

neighbourhood of any given period- n point tends to 1. If this period- n point is stable, once the orbit enters this neighbourhood it will converge to the periodic orbit of size n and thus be asymptotically not transitive or diverging almost surely.

However, the third defining feature for a map to be chaotic in terms of the definition by Devaney [1986] is the existence of a dense set of periodic points. This means for any subinterval $B \subseteq J$ there exists at least one periodic point. Taking B to be arbitrarily small, this implies the existence of infinitely many periodic points, all of which must be either unstable or neutral for all finite period lengths. It should be noted that the number of periodic points, although infinite, is countable whilst the number of points in J is uncountable, therefore the probability of choosing a periodic point at random from a uniform probability distribution over J is still zero. Periodic orbits are thus atypical for chaotic maps.

For one dimensional maps, Sarkovskii [1964] gives the following Theorem which can help determine which period lengths are present in a map.

Theorem 3.1.2. *Consider the following ordering of all positive integers*

$$3, 5, 7, \dots, 2 \times 3, 2 \times 5, 2 \times 7, \dots, 2^2 \times 3, 2^2 \times 5, 2^2 \times 7, \dots, 2^3, 2^2, 2^1, 1$$

that is all odd numbers except 1, followed by two times all odd numbers except 1, followed by two squared times all odd numbers except 1 and so on until 2^∞ times all odd numbers except 1 and then all the powers of 2 in decreasing order until 1.

If a one dimensional map, $\zeta : J \rightarrow J$, has a periodic orbit with period p_1 then, for any p_2 such that p_2 appears after p_1 in the ordering of positive integers considered the map has a periodic orbit with period p_2 .

This theorem implies if the one dimensional map $\zeta : J \rightarrow J$ has a periodic orbit of period p where p is not a power of 2 then the map has an infinite number of periodic orbits, including all orbits of periods which are powers of 2. For example, if a map has an orbit of period 3 it must have orbits of all other possible period lengths. In addition, since one dimensional chaotic maps, by the definition of Devaney [1986], must have infinitely many periodic points, they must include all orbits of periods which are powers of 2.

Work done by Li and Yorke [1975] shows the existence of a period 3 orbit, in addition to implying all other period length orbits exist, also implies the existence of an uncountable set of orbits which remain nonperiodic. Li and Yorke [1975] describe this situation as 'chaos'. If the period 3 orbit is stable, these nonperiodic orbits still exist, but they are unstable and the probability of choosing a point on a nonperiodic orbit at random from a uniform distribution over J is zero. Hence, starting at a typical initial condition for a map with a stable period 3 orbit will produce an orbit which converges to a stable state and the map is thus not typically chaotic.

3.1.4 Invariant Density

This section introduces some additional features of some chaotic maps which will be of use during the discussion of the chaotic intermittency maps which will be introduced in section 3.2, namely *ergodicity*, *invariant measures* and *invariant densities*.

Consider a map $\zeta : J \rightarrow J$. Take an infinite number of initial conditions randomly from J with a smooth density $\chi_0(w)$ such that $\int_a^b \chi_0(w)dw$ is the fraction of initial conditions in the interval $[a, b]$. Applying the map ζ to each initial condition produces a new set of values with a new density, $\chi_1(w)$. Further iterations of the map produce densities $\chi_2(w)$, $\chi_3(w)$, and so on. An *invariant density*, $\chi(w)$, for a map, ζ , is a density such that

$$\chi_{t+1}(w) = \chi_t(w) = \chi(w),$$

that is, a density which is invariant to iterations of the map. The existence of an invariant density is assumed for the remainder of this section. Discussion of when an invariant density exists can be found in Sinai [1972] and Bowen and Ruelle [1975].

The invariant density of a map, ζ , can be used to find *invariant measures*, that is measures of a map which remain constant over iterations. For example, let $\mu(w)$ be the invariant measure defined by

$$\mu(w) = \int_J w\chi(w)dw, \quad (3.6)$$

that is $\mu(w)$ is the mean of the infinitely many points taken randomly from the invariant density. Other invariant measures include

$$\mu(w^2) = \int_J w^2\chi(w)dw,$$

and

$$\mu(w\zeta^n(w)) = \int_J w\zeta^n(w)\chi(w)dw, \quad (3.7)$$

which represent the mean of each of the infinite points taken randomly from $\chi(w)$ squared and the mean of each point multiplied by the point produced after n iterations of the map ζ . In general, for any *Hölder continuous function*, that is a function, $\phi(w)$, for which there exists a $\gamma \in (0, 1)$ and a $C < \infty$, independent of x, y , such that

$$|\phi(x) - \phi(y)| \leq C|x - y|^\gamma, \quad (\text{for all})x, y \in D$$

where D is the domain of the function, then

$$\mu(\phi(w)) = \int_J \phi(w)\chi(w)dw. \quad (3.8)$$

A useful mathematical tool is the *Perron-Frobenius* operator, $(P_\zeta\phi)(w)$, may be defined as

$$(P_\zeta\phi)(w) = \sum_{t:\zeta(t)=w} \frac{\phi(t)}{|\zeta'(t)|},$$

although a more general definition can be seen in Baladi [2000]. The invariant density, $\chi(w)$, of a map ζ , is a fixed point of this operator such that

$$(P_\zeta\chi)(w) = \sum_{t:\zeta(t)=w} \frac{\chi(t)}{|\zeta'(t)|} = \chi(w),$$

as shown by Lawrance [2001]. Also, see Lawrance et al [1995],

$$R_{\phi,\psi}(u) \leq \left(\max_J |\phi(w)| \right) \left\{ \left\| (P_{\psi\zeta}^u\chi)(w) - \chi(w) \int_J \psi(t)dt \right\|_1 \right\},$$

where $R_{\phi,\psi}(u)$ is the covariance between $\phi(\zeta^u(w))$ and $\psi(w)$. Since $\psi(w)$ can be normalised such that $\int_J \psi(t)dt = 1$, the rate at which $\left\| (P_{\psi\zeta}^u\chi)(w) - \chi(w) \right\|_1$ tends to zero gives an upper bound on the rate of decay of the correlations.

Now, thus far the invariant measures have been defined in terms of infinitely many points taken randomly and independently from the invariant density. In this setting, it can be seen that the invariant density, $\chi(w)$, for a chaotic map is a probability density of these points. In particular the mean, variance and autocovariance of the infinitely many points taken randomly from $\chi(w)$ can be defined as $\mu(w)$, $\mu(w^2) - \mu(w)^2$ and $\mu(w\zeta^n(w)) - \mu(w)^2$, respectively. Consider now an orbit $\{w_t\}$, $t \in \mathbb{N}$, generated by repeated iterations of a map ζ from a single initial point w_0 . Other than the choice of initial value w_0 , the orbits are completely deterministic and use of stochastic methods has been debated, see for example Berliner [1992b]. However, due to the difficulties of long term prediction discussed in section 3.1.1, probabilities, such as $P(w_n < 0.5|w_0) = p$ for $p \in (0, 1)$ and large n , may intuitively seem to make sense even when the initial value, w_0 , is known.

Now, if the map is *ergodic* and w_0 is a typical value taken from the invariant density, then, for an invariant measure $\mu(\phi(w))$ defined by equation 3.8,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^n \phi(w_t) = \mu(\phi(w)),$$

and

$$\lim_{n \rightarrow \infty} \frac{1}{n} (\#\{k \in [0, n-1] : \zeta^k(w) \in A\}) = \int_A \chi(w) dw, \quad \text{for } A \subset J.$$

Hence, for an ergodic map the mean, variance and autocovariances of each orbit simulated from a typical initial value w_0 tends to the true mean, variance and autocovariances of the map, as $n \rightarrow \infty$. Ergodicity also gives meaning to probability, in a relative frequency sense, to be assigned to the event that a typical orbit of length n visits a set $A \subset J$.

This also implies that if initial values are taken from a smooth initial density $\chi_0(w) \neq \chi(w)$, such that all the initial values are within J , then, for an ergodic map, $\chi_n(w) \rightarrow \chi(w)$ as $n \rightarrow \infty$. In addition, if the map, ζ , is ergodic and the initial condition w_0 is taken randomly from the invariant density, the orbit $\{w_t\}$, $t \in \mathbb{N}$, is a stationary process, see, for example, Berliner [1992a]. For further discussion on ergodicity. see, for example, Walters [1975].

3.2 Chaotic Intermittency Maps

The general properties of chaotic maps given in section 3.1 highlight the possibility of defining maps which produce orbits with long-memory characteristics. Section 3.1.4 stated, with reference to Berliner [1992a], if the map, ζ , is ergodic and an invariant density, $\chi(w)$, exists, an orbit $\{w_t\}$, $t \in \mathbb{N}$, with initial condition, w_0 , taken randomly from $\chi(w)$, is a stationary process. Section 3.1.3 discussed how the existence of neutral fixed points in the map can lead to longer escape/return times and Theorem 3.1.1, given in section 3.1.2, relates these return times to the possibility of subexponential decay of the orbits ACF.

The class of chaotic intermittency maps introduced in this section were first studied by Manneville and Pomeau [1980], motivated by the need to model series exhibiting long periods of laminar behaviour with short bursts of erratic behaviour. These maps have seen much development since then, see for example Bhansali et al [2005], [2006] and Bhansali and Holland [2008b]. Each of the maps have at least one neutral fixed point, are ergodic and admit an invariant density. The orbits produced by these maps are thus stationary and possess ACFs with subexponential decay rates. The majority of long-memory models discussed in the literature are largely based around the concept that fractionally differencing produces a linear process with short memory. The methods of estimating the long-memory parameter are also often based around this assumption.

The set of chaotic intermittency maps provide a new alternative method of generating long memory time series. They are non-linear and yet can produce stationary processes with long-memory properties. Furthermore, chaotic intermittency

maps have been found whose long-memory parameters lie on the boundaries of long-memory, yet still remain stationary. This is discussed further in Chapter 5.

The four main chaotic intermittency maps brought together in this study are the asymmetric cusp map, the symmetric cusp map, the polynomial map and the logarithmic map. These are defined in sections 3.2.1 - 3.2.3, where some generalizations of these maps are also discussed.

3.2.1 The Cusp Maps

The class of cusp maps have been investigated previously by Balakrishnan et al [1997], [2001], [2001]. Bhansali and Holland [2008b] define the family of *extended symmetric cusp maps* by

$$\zeta_{\tau,\theta}(w) = \begin{cases} w + (w + 1)^{1+\tau}, & -1 \leq w \leq -1 + \delta, \\ \tilde{\zeta}(w), & -1 + \delta < w \leq -\delta, \\ 1 - |w|^\theta, & -\delta < w \leq 0, \\ \zeta_{\tau,\theta}(-w), & 0 < w \leq 1, \end{cases}$$

where $\tau > 0$, $\theta \in (0, 1)$, $\delta > 0$ is a small constant and $\tilde{\zeta}(w)$ is chosen such that the map shows sensitivity dependence on $[-1 + \delta, -\delta]$ and such that $\zeta_{\tau,\theta}(w)$ maps J on to itself where $J = [-1, 1]$. The derivative of these maps near $w = -1$ is

$$\zeta'_{\tau,\theta}(w) = 1 + (1 + \tau)(w + 1)^\tau \quad \text{for } -1 \leq w \leq -1 + \delta,$$

therefore $w = -1$ is a neutral fixed point and the maps have a laminar region near this point and a chaotic region away from it, leading to intermittency.

Theorem 3.2 of Bhansali and Holland [2008b] gives the following result for the rate of decay of correlations for orbits from extended symmetric cusp maps.

Theorem 3.2.1. *For the family of extended symmetric cusp maps with $\tau\theta < 1$, there exists a constant $C_{\tau,\theta}$ such that the ACF, $R(u)$, of typical orbits is of the form*

$$R(u) \sim C_{\tau,\theta} u^{1-1/(\tau\theta)}$$

This family of maps is a generalization of the symmetric cusp map, which is defined as,

$$\zeta_S(w) = 1 - 2\sqrt{|w|}$$

over $J = [-1, 1]$. For the symmetric cusp map, Bhansali and Holland [2008b] prove the covariances of the symmetric cusp map, $R_S(u)$, are of the form

$$R_S(u) = \frac{4}{9u} + o\left(\frac{1}{u}\right) \quad \text{as } u \rightarrow \infty, \quad (3.9)$$

which is in agreement with Theorem 3.2.1 when $\tau\theta = 0.5$. Balakrishnan et al [1997] gives the invariant density of the symmetric cusp map, $\chi_S(w)$, over $J = [-1, 1]$ as

$$\chi_S(w) = \frac{1-w}{2}, \quad (3.10)$$

and zero everywhere else. Substitution of this into equations 3.6-3.7 can give explicit forms for the mean, variance and covariance at lag one, namely $E(w_t) = -1/3$, $Var(w_t) = 2/9$ and $R_S(1) = 8/63$ for $t \in \mathbb{N}$.

Theorem 4.1 of Bhansali and Holland [2008b] shows the spectral density of a typical orbit from the symmetric cusp map, as the orbit length $n \rightarrow \infty$, is of the form

$$f_S(\lambda) = G \log \left(\frac{1}{2 \sin(\frac{\lambda}{2})} \right) \quad \text{as } \lambda \rightarrow 0,$$

for a positive bounded constant G . Note, since $\sin(x) \rightarrow x$ as $x \rightarrow 0$, this is equivalent to

$$f_S(\lambda) = G \log \left(\frac{1}{\lambda} \right) \quad \text{as } \lambda \rightarrow 0. \quad (3.11)$$

Now, as discussed in section 2, the long memory parameter, d , of a process x_t is defined such that the ACF, $R_x(u)$, is of the form

$$R_x(u) \sim B(u)u^{2d-1} \quad \text{as } u \rightarrow \infty, \quad (3.12)$$

and the spectral density, $f_x(\lambda)$, is of the form

$$f_x(\lambda) = G(\lambda)\lambda^{-2d} \quad \text{as } \lambda \rightarrow 0, \quad (3.13)$$

where $B(u)$ and $G(\lambda)$ are constant or slowly varying functions. Comparison of equations 3.12 and 3.13 with equations 3.9 and 3.11 shows the asymptotic value of the long memory parameter from typical orbits for the symmetric cusp map is given by $d = 0$. A value of $d = 0$ is usually associated with short memory, however the correlations still decay at a subexponential rate, specifically a harmonic rate, and the spectral density is still unbounded at zero. Hence, despite $d = 0$, the orbits still possess properties of long memory and can thus be considered to be on the lower boundary of long memory.

A related family of maps is the family of *extended asymmetric cusp maps*, defined by

$$\zeta_{\tau,\theta}(w) = \begin{cases} w + (w+1)^{1+\tau}, & -1 \leq w \leq -1 + \delta, \\ \tilde{\zeta}(w), & -1 + \delta < w \leq -\delta, \\ 1 - |w|^\theta, & -\delta < w \leq 0, \\ -\zeta_{\tau,\theta}(-w), & 0 < w \leq 1, \end{cases}$$

where $\tau > 0$, $\theta \in (0, 1)$, $\delta > 0$ is a small constant and $\tilde{\zeta}(w)$ is chosen such that the map shows sensitivity dependence on $[-1 + \delta, -\delta]$ and such that $\zeta_{\tau,\theta}(w)$ maps J on to itself where $J = [-1, 1]$. This family of maps differs by the negative sign for $0 < w \leq 1$ which makes the maps asymmetric and produces another neutral fixed point at $w = 1$. They are a generalization of the asymmetric cusp map, $\zeta_A(w)$, which is defined over $J = [-1, 1]$ as:

$$\zeta_A(w) = \begin{cases} 1 - 2\sqrt{-w} & \text{if } -1 \leq w \leq 0, \\ 2\sqrt{w} - 1 & \text{if } 0 < w \leq 1. \end{cases} \quad (3.14)$$

An alternative generalization of the asymmetric cusp map, given by Bhansali et al [2005], is the maps $\zeta_\theta(w)$ such that

$$\zeta_\theta(w) = \begin{cases} \theta - 2\sqrt{-w}, & -1 \leq w \leq -1/2, \\ \theta - 2\sqrt{-w} + 8(\theta - 1)(1/2 + w)^3, & -1/2 < w \leq 0, \\ 2\sqrt{-w} - \theta + 8(\theta - 1)(1/2 - w)^3, & 0 < w \leq 1/2, \\ 2\sqrt{-w} - \theta, & 1/2 < w \leq 1, \end{cases}$$

where $\theta \in [1, 16/15)$ and the map is defined over $J = [-1, 1]$. When $\theta = 1$ the asymmetric map defined in equation 3.14 is produced.

The asymmetric map has a uniform invariant density, $\chi_A(w)$, such that

$$\chi_A(w) = \frac{1}{2}, \quad (3.15)$$

over the full range of $J = [-1, 1]$ and zero elsewhere. Substitution of this into equations 3.6-3.7 can give explicit forms for the mean, variance and covariance at lag one, namely $E(w_t) = 0$, $Var(w_t) = 1/3$ and $R_A(1) = 3/5$ for $t \in \mathbb{N}$. Bhansali et al [2005], [2006] hypothesise that the ACF of typical orbits from the asymmetric map, $R_A(u)$ are of the form

$$R_A(u) = O\left(\frac{1}{u}\right) \quad \text{as } u \rightarrow \infty,$$

that is, they decay at a harmonic rate as with those for the symmetric cusp map. Theorem 4.1 of Bhansali and Holland [2008b] is once again applicable to show the spectral density of a typical orbit from the asymmetric cusp map, as the orbit length

$n \rightarrow \infty$, is once again of the form given in equation 3.11. As with the symmetric cusp map, the asymptotic value of the long memory parameter is $d = 0$ although the orbits still possess long memory properties.

3.2.2 The Polynomial Maps

The polynomial map, see Liverani et al [1999] and Young [1999], is defined over the range $J = [0, 1]$ by

$$\zeta_\alpha(w) = \begin{cases} w(1 + 2^\alpha w^\alpha) & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

Where $\alpha > 0$ is a parameter of the map.

When $\alpha \in (0, 1)$ the invariant density, see Thaler [1980], $\chi_\alpha(w)$ over J is known to be of the form

$$\chi_\alpha(w) = \frac{V_\alpha(w)}{w^\alpha} \quad (3.16)$$

and zero elsewhere, where $V_\alpha(w)$ depends on the value of α and for each fixed α it is a piecewise continuous, uniformly bounded function which is also bounded away from zero. An explicit expression for $V_\alpha(w)$, and therefore $\chi_\alpha(w)$ is unknown. This implies the mean and variance cannot be found as they were with the cusp maps.

For the polynomial map, Bhansali and Holland [2008b], with reference to results given by Sarig [2002] and Gouëzel [2004b], give the correlations for typical orbits, $r_\alpha(u)$, to be of the form

$$\lim_{u \rightarrow \infty} u^{(1/\alpha)-1} r_\alpha(u) = N, \quad (3.17)$$

for some finite constant N . Hence the correlations of typical orbits decay at a polynomial rate and the spectral densities as the orbit lengths $n \rightarrow \infty$, for $\alpha \neq 0.5$, are of the form

$$f_\alpha \sim G(\lambda) \lambda^{(1/\alpha)-2} \quad \text{as } \lambda \rightarrow 0, \quad (3.18)$$

where $0 < G(\lambda) < \infty$ is a constant or slowly varying function as $\lambda \rightarrow 0$. For the case when $\alpha = 0.5$, equation 3.17 shows the correlations decay at a harmonic rate and thus Theorem 4.1 of Bhansali and Holland [2008b] is once again applicable to show the spectral density is of the same form as for the cusp maps given in equation 3.11. Comparison of equations 3.17 and 3.18 with equations 3.9 and 3.11 shows the asymptotic value of the long memory parameter from typical orbits for the polynomial map is given by

$$d = 1 - \frac{1}{2\alpha}.$$

When $\alpha \in (0, 1/3]$ this gives the corresponding asymptotic value of $d \leq -0.5$. The correlations are thus asymptotically like a fractionally differenced process which has been differenced a further $p = [1 - 1/2\alpha]$ times, where $[x]$ is the integer part of x . For $\alpha \in (1/3, 0.5)$ this gives the corresponding asymptotic value of $d \in (-0.5, 0)$ and thus the orbits admit intermediate memory. For $\alpha \in (0.5, 1)$ this gives the corresponding asymptotic value of $d \in (0, 0.5)$ and thus the orbits admit long memory. For the special case of $\alpha = 0.5$, the asymptotic value of d is zero. However, as with the cusp maps mentioned in section 3.2.1, although $d = 0$, the correlations still decay at a subexponential rate and the spectral density has a singularity at zero, hence the orbits still displays long memory properties and can be considered to lie on the lower boundary of long memory.

3.2.3 The Logarithmic Maps

The logarithmic map was introduced by Holland [2005] as an extension of the polynomial map. It is defined over the range $J = [0, 1]$ by

$$\zeta_\beta(w) = \begin{cases} w[1 + 2(\log 2)^{-(1+\beta)}w(-\log w)^{1+\beta}] & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

Where β is a parameter of the map. The multiplying function $2(\log 2)^{-(1+\beta)}$ is chosen to ensure that $\lim_{w \rightarrow 1/2^-} \zeta_\beta(w) = 1$.

As with the polynomial map, the exact form of the invariant density, $\chi_\beta(w)$, for the logarithmic map is currently unknown, however it is known, see Holland [2005], that the invariant density over J for $\beta \in (0, 0.5)$ is of the form

$$\chi_\beta(w) = \frac{V_\beta(w)}{w \log(1/w)^{\beta+1}} \quad (3.19)$$

and zero elsewhere, where $V_\beta(w)$ depends on the value of β and for each fixed β it is a piecewise continuous, uniformly bounded function which is also bounded away from zero. For $\beta \in [0.5, 1]$ the invariant density is a bounded piecewise continuous function.

The logarithmic map can be combined with the polynomial map, see Holland [2005], to give the family of *generalized polynomial-logarithmic maps* defined over the range $J = [0, 1]$ by

$$\zeta_{\alpha,\beta}(w) = \begin{cases} w \left[1 + (2w)^\alpha \frac{L_\beta(w)}{L_\beta(0.5)} \right] & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1, \end{cases} \quad (3.20)$$

where $L_\beta(w)$ is a slowly varying function dependent on a parameter β which is twice differentiable.

Theorem 3.3 of Bhansali and Holland [2008b] , making use of Gouëzel [2004a], gives the covariances, $R_\beta(u)$, of a typical orbit from the logarithmic map to be of the form

$$R_\beta(u) = \frac{\mu_\beta \chi_\beta(1/2)}{2\beta(\log u)^\beta} + o\left(\frac{1}{(\log u)^\beta}\right) \quad \text{as } u \rightarrow \infty, \quad (3.21)$$

where $\mu_\beta = E(w_t)$ for the logarithmic map and the covariances of the generalized polynomial-logarithmic map, $R_{\alpha,\beta}(u)$ to be of the form

$$R_{\alpha,\beta}(u) = \nu_{\alpha,\beta} L_\beta^*(u) u^{1-1/\alpha} + o\left(L_\beta^*(u) u^{1-1/\alpha}\right) \quad \text{as } u \rightarrow \infty, \quad (3.22)$$

where $\nu_{\alpha,\beta}$ is a constant dependent on α and β and L_β^* is a slowly varying function dependent on β .

For the logarithmic map, Theorem 4.2 of Bhansali and Holland [2008b] is applicable which gives the spectral density, $f_\beta(\lambda)$, of a typical orbit with length $n \rightarrow \infty$ of the form

$$f_\beta \sim G(\lambda) \lambda^{-1} \left(\log\left(\frac{1}{\lambda}\right)\right)^{-1-\beta} \quad \text{as } \lambda \rightarrow 0, \quad (3.23)$$

where $G(\lambda)$ is a constant or bounded continuous function.

Comparison of equations 3.21 and 3.23 with equations 3.9 and 3.11 shows the asymptotic value of the long memory parameter from typical orbits for the logarithmic map is given by $d = 0.5$. A value of $d = 0.5$ is usually considered to be non-stationary, however, for $\beta \in (0, 2\log(2) - 1)$, the map is ergodic and admits an invariant density and thus typical orbits are stationary . Hence the orbits produced by the logarithmic map can thus be considered to be on the upper boundary of long memory.

For the generalized polynomial-logarithmic map, Theorem 4.3 of Bhansali and Holland [2008b] is applicable which gives the spectral density, $f_{\alpha,\beta}(\lambda)$, of a typical orbit with length $n \rightarrow \infty$ of the form

$$f_{\alpha,\beta} \sim G_\beta\left(\frac{1}{\lambda}\right) \lambda^{(1/\alpha)-2} \quad \text{as } \lambda \rightarrow 0, \quad (3.24)$$

where $G_\beta(\lambda)$ is a non-bounded slowly varying function at infinity dependent on β .

Comparison of equations 3.22 and 3.24 with equations 3.9 and 3.11 shows that, as with the polynomial map, the asymptotic value of the long memory parameter from typical orbits for the generalized polynomial-logarithmic map is given by $d = 1 - 1/(2\alpha)$. When $\alpha = 0.5$, the map produces orbits with asymptotic d values of zero,

whilst when $\alpha = 1$ the map produces orbits with asymptotic d values of 0.5. The behaviour at these points is then dependent on the choice of $L_\beta(w)$ and the value of β . When $d = 0$, the orbits could display short memory, boundary behaviour of long memory or even boundary behaviour of intermediate memory. When $d = 0.5$, the orbits could be stationary or non-stationary.

4. ESTIMATION OF THE CORRELATION DECAY RATES FOR CHAOTIC INTERMITTENCY MAPS

4.1 Simulation Study

The simulation study carried out here examines the question of estimating, in a statistical sense, the rate of decay of the correlations for the chaotic intermittency maps introduced in section 3.2 when only a finite series length, T , of the orbits produced by the maps are available and the generating maps is treated as unknown. The situation in which some information regarding the map is considered to be known is explored further in chapter 6.

A range of methods were used in estimation, chosen to cover the three approaches of estimation for the long memory parameter of non-parametric, semi-parametric and parametric and hence provide comparison between the three. For the non-parametric estimates, the local Whittle and GPII methods are implemented, for the semi-parametric approach, estimates are found using the FAR and FExp methods and finally the FARIMA method is used to find parametric estimates. An introduction to each of these methods was given in chapter 2, whilst further details of application of these methods here can be found in section 4.1.1.

The motivation behind this study has two main goals. First, the empirical properties of these estimation techniques are often studied in the literature using series generated from linear long memory models, see for example Taqqu and Teverovsky [1996], and many of the methods of estimating the long memory parameter assume both linearity and sometimes Gaussianity of the series under investigation, see chapter 2 for a review. Non-linear examples of long memory processes are usually the result of a transformation applied to linear process, see, for example, Giraitis and Taqqu [1999], Palma and Zavallos [2004] and Palma [2007]. The chaotic intermittency maps introduced in section 3.2 thus give an important alternative to the majority of long memory processes previously studied, in that they exhibit examples of non-linear, non-Gaussian time series which are the result of the intermittency rather than a transformation. The polynomial map with $\alpha \neq 0.5$, in particular, has spectral densities near zero of the standard form assumed by all of the estimation methods considered and the results should, therefore, throw light on the robustness of these estimation techniques in this non-standard case.

In addition, the orbits of the cusp maps and the polynomial map with $\alpha = 0.5$ show boundary long memory behaviour for which a long memory parameter of $d = 0$ can still show signs of 'weak' long memory and the logarithmic maps have a value of

$d = 0.5$ yet are still stationary. Long memory models such as FARIMA(p, d, q) and FExp(s) models can not account for this boundary behaviour, and thus examples of this behaviour in the literature seem quite limited. The results should, therefore, provide information about the empirical behaviour of these estimation techniques when the long memory parameter takes these boundary values. Some specific current examples of boundary behaviour can be found in Martin and Eccleston [1992], Martin and Walker [1997] and Palma [2007] page 61, though a general model has not previously been defined and the effect on the sampling properties of estimates has not been discussed. This is explored in greater detail in Chapter 5, where the sampling properties of the GPII and Local Whittle methods in such situations are derived and a new extension to the FARIMA model to include these boundary conditions is defined.

These particular methods have been chosen as the finite sample properties of the estimates produced in standard situations have been extensively studied in the literature by various authors.

Robinson [1995b] compares the two non-parametric methods for time series of length $n = 64$, $n = 128$ and $n = 256$, finding the GPII estimates to be generally less biased whilst the local Whittle estimates have smaller variance. Also noted is the biases of the local Whittle estimates tend to be negative, whilst the estimates of the GPII methods tend to have negative bias for smaller values of d_0 and positive bias for larger values of d_0 .

Taqqu and Teverovsky [1997] investigate the finite sampling properties of the local Whittle estimates on simulated FARIMA(p, d_0, q) series, for $p, q \in (0, 1)$. They find for series lengths of $n = 100$ the sample standard deviations are much larger than the asymptotic results and indeed so large that the method is deemed unsuitable. However, for series length $n = 10^4$ they find the theoretical asymptotic results appear to hold very well in practise and that provided m is chosen correctly the local Whittle method seems robust to changes in the short memory component. It is also suggested that the asymptotic results may hold reasonably well for a series length of $n = 10^3$ although the results are not presented.

Hurvich and Beltra [1994] carry out simulation studies for the performance of the GPII method on simulated FARIMA($1, d_0, 0$) series of lengths $n \in (500, 1000, 2000, 3000)$ using various values of l and m . They find the choice of $l = 1$ to be superior to $l > 1$. They also report the biases and MSEs of the estimates tend to be small provided the AR parameter is not close to unity.

Other simulation studies by Bhansali and Kokoszka [2001a], Taqqu and Teverovsky [1996] and Taqqu, Teverovsky and Willinger [1995], amongst others, show the asymptotic results for the GPII and local Whittle methods appear to hold for time series of lengths $n \geq 1000$ generated from linear Gaussian processes. Bardet et al. [2003] also give empirical evidence to suggest the GPII and local Whittle methods may all be robust to estimating d_0 from a FARIMA($0, d_0, 0$) process with non-Gaussian

innovations.

For the parametric FARIMA method, studies by Taqqu and Teverovsky [1996], Bisaglia and Bordignon [2002] and Smith et al. [1997] have shown that when the model is correctly specified, the parametric approach outperforms the non-parametric and semi-parametric approaches, though when the 'true' model is misspecified or does not exist this method can lead to large bias and/or variance.

Results by Bhansali and Kokoszka [2001a] show heavy tailed and bimodal distributions begin to appear when estimating the long memory parameter using the FARIMA method from an α -stable FARIMA process, that is a FARIMA process with innovations, ε_t , such that

$$P(|\varepsilon_t| > x) = x^{-\alpha}L(x),$$

for some slowly varying function $L(x)$ and

$$P(\varepsilon_t > x)/P(|\varepsilon_t| > x) \rightarrow a, \quad P(\varepsilon_t < -x)/P(|\varepsilon_t| > x) \rightarrow b,$$

as $x \rightarrow \infty$, where a and b are nonnegative numbers satisfying $a + b = 1$. They thus warn against naive use of the FARIMA method when 'outliers' may be present, although a theoretical explanation as to why this occurs is not given. For the FAR method, Bhansali and Kokoszka [2001a] prove the estimates are still asymptotically consistent for α -stable FARIMA process with $\alpha \in (1, 2)$ and their empirical results support this finding. It is also seen that the FAR and FExp methods are more robust to model changes than a parametric approach with smaller variances than the non-parametric approach.

A second motivation for this simulation study is to explore the extent to which the asymptotic rates of correlation decay of the chaotic intermittency maps remain valid for finite values of T , since the rate of convergence to these asymptotic results is currently unknown.

Previous simulation studies involving the chaotic intermittency maps discussed in section 3.2 can be found in Bhansali, Holland and Kokoska [2006] and Bhansali and Holland [2008b]. The simulation study of Bhansali, Holland and Kokoska [2006] begins with estimates of the invariant densities of the polynomial and logarithmic maps with various values of α and β respectively. Since explicit theoretical forms for these invariant densities are unknown, it is not possible to quantify how well the simulations fit. However, in a neighbourhood of $w = 0$, it is noted that the estimated invariant densities for the polynomial map take values close to $w^{-\alpha}$ and the estimated invariant densities for the logarithmic map take values close to $w^{-1} \log(1/w)^{-1-\beta}$ and thus their shapes accord with the asymptotic results given by equations 3.16 and 3.19.

Bhansali, Holland and Kokoska [2006] then continue their simulation study, presenting results on correlations, partial correlations and the autoregressive order selected for the polynomial and logarithmic maps with various values of α and β ,

respectively, and also the asymmetric cusp map. They find the overall tendencies generally agree with the asymptotic results. For the polynomial map it is found, particularly for $\alpha > 0.5$, that the correlations tend to be positive for all lags up to $u = 100$ and that the rate of decay of correlations decreases as α increases. Additionally, the mean AR order selected is clearly shown to increase with α , with the change due to α being most noticeable as α increases from 0.55 to 0.8. The distributions of estimates of the means and correlations seem to not be Normal, particularly for $\alpha > 0.5$, and this deviation from Normality increases with α . Similar results are reported for the logarithmic map, with the memory seeming to decrease with β , though the effect of β being less noticeable than that of α for the polynomial map. For the asymmetric cusp map, the correlations, although generally larger for smaller lags, appear to decay more rapidly than the logarithmic map and polynomial map with $\alpha > 0.5$, though still remain significantly positive at large lags.

Although the general patterns seem to agree with the asymptotic results, Bhansali, HOLLAND and Kokoska [2006] also show that attempts to estimate α for the polynomial map from regression of $\log(R(u))$ onto $\log(u)$ and β for the logarithmic map from regression of $\log(R(u))$ onto $\log(\log(u))$ provides somewhat poor approximations to the actual values of α and β and varies considerably depending on how many initial values of the covariances are ignored.

Bhansali and HOLLAND [2008b] give simulation results on the symmetric cusp map and further results on the polynomial and logarithmic map in the frequency domain. For the symmetric cusp map, the estimated invariant density generally accords with the theoretical triangular density, though the relative frequency of the orbit remaining close to the neutral fixed point is underestimated, whilst the relative frequencies away from the neutral fixed point are overestimated. The correlations are shown to remain significantly positive for large lags and appear similar to those of the asymmetric cusp map reported in Bhansali, HOLLAND and Kokoska [2006]. The simulated results in the frequency domain, for all three maps, only partially agrees with the asymptotic theory. The results for the symmetric cusp map show if d is estimated by a log-periodogram regression, this estimate could have strong positive bias. For the polynomial map with $\alpha < 0.5$ evidence of intermediate memory is not seen and the behaviour of the orbits appears less sensitive to the choice of α than the asymptotic theory implies. Conversely, estimates of d for the logarithmic map suggest the behaviour of the orbits appears more sensitive to the choice of β than the asymptotic theory implies, with the agreement between the theoretical results and the simulated results decreasing as β increases.

The simulation study presented here builds on the works of Bhansali, HOLLAND and Kokoska [2006] and Bhansali and HOLLAND [2008b], focusing now on estimation of the long memory parameter, d . The study looks at the polynomial map, the logarithmic map, the symmetric cusp map and the asymmetric cusp map, discussed in section 3.2. The studies by Bhansali, HOLLAND and Kokoska [2006] and Bhansali and HOLLAND [2008b] have suggested some difficulties in estimating d may occur due

to a lack of convergence of the properties of the simulated orbits to the asymptotic results. Further simulation studies are carried out in section 4.2 which attempt to resolve some of these issues.

4.1.1 Plan of Study

The computational set up for this simulation study follows closely from the simulation studies suggested in Bhansali, Holland and Kokoska [2006] and Bhansali and Holland [2008b], which also study the chaotic intermittency maps described in section 3.2.

For the asymmetric cusp and symmetric cusp maps the invariant densities are known, see equation 3.15 and equation 3.10 respectively. The initial values for all simulated orbits of these maps were taken randomly from these distributions. Since the invariant densities are not known for the polynomial and logarithmic maps, the initial values were generated instead via an exponential distribution with mean 0.2, truncated at 1, as suggested in Bhansali, Holland and Kokoska [2006].

For each simulated orbit from each map a stretch of $N = 10^7$ iterations were generated from these initial values, but only the last $T = 10^4$ observations, $\{w_t, t = M + 1, \dots, M + T\}$, $M = 10^7 - 10^4$, were retained. This 'burn in' time, as suggested in Bhansali, Holland and Kokoska [2006], helps to avoid possible 'transient' effects. The polynomial map was generated using six different values for alpha, namely, $\alpha = 0.3, 0.45, 0.5, 0.65, 0.8$ and 0.9 . The logarithmic map was generated with four different values for beta, $\beta = 0.05, 0.15, 0.25$ and 0.3 . More values of α were considered than β since the theoretical value of d for the polynomial map is related to α . For each of the maps, 1000 orbits were simulated in the above manner.

For each simulated orbit, the local Whittle, GPII, FARIMA, FAR and FExp methods of long memory parameter estimation were applied to find estimates of the long memory parameter d . These estimated \hat{d} for each method were then retained for study. The following choices were made for each method, see section 2.3 for further details.

For the local Whittle method, m was chosen as 155. This was based on studies by Bhansali and Kokoszka [2001a] and Taqqu and Teverovsky [1996], which suggest m be taken in the range of $1/50$ to $1/20$ of all frequencies and show that usage of approximately the first $1/32$ of the spectrum tends to give reliable results. A brief simulation study carried out here showed the estimated values of d remained essentially the same over a similar range for the chaotic intermittency maps.

The GPII method was implemented with $m=100$. This is equal to $T^{0.5}$ which was suggested by Geweke and Porter-Hudack [1983] and is within the range considered by Crato and Lima [1994] and Porter-Hudak [1990]. As with the local Whittle method, brief simulation studies suggested little change with the GPII estimates over the range of $1/50$ to $1/20$ of all frequencies.

For the FARIMA method, the Whittle likelihood estimates were estimated for all FARIMA(p,d,q) models with $p, q \in (0, \dots, 5)$. The estimates of d from the models chosen by minimisation of AIC and BIC were both then retained. The order chosen by each selection criteria and the estimated short memory co-efficients for each were also retained. The exact likelihood estimates were not obtained due to computational time.

Whittle likelihood estimates were also used for the FAR method, which was implemented over the range $p = 0, \dots, 20$ and again the estimates of d from the models chosen by minimisation of AIC and BIC, along with the order chosen by each selection criteria and the estimated short memory co-efficients for each, were then retained.

The semiparametric FEXP method was used with $m = 10$ and $h_j(x) = \cos(jx)$. The choice of s was found using Mallows's C criterion over the range of $s = 0, \dots, 20$. These choices are based on the method given by Moulines and Soulier [1998]. Simulation studies showed the estimates were not very sensitive to the choice of m . Other forms of $h_j(x)$ were not considered. The estimates of d and the selected choice of s , along with the estimated short memory coefficients for the selected model, were recorded for each simulation.

4.1.2 Simulation Results

The histograms of the estimates of d for the FARIMA method using AIC and BIC are given in figures 4.1 and 4.2 respectively. As previously mentioned, a parametric approach can lead to poor estimates when the assumed underlying model is misspecified and this is reflected here. The results seem comparable to those found by Bhansali and Kokoszka [2001a] when the method is applied to α -stable FARIMA processes and add as an additional warning against naive use of these methods.

The estimates found using AIC show a general lack of convergence to a single value for d , with values covering the whole range of $\hat{d} \in (-0.5, 0.5)$ for all the maps. It can be seen to best converge for the polynomial map with $\alpha = 0.3$, for which it centres around $\hat{d} = 0$. The situation gradually deteriorates as α increases, with the distributions for $\alpha \geq 0.65$ appearing somewhat bimodal. This suggests that the lack of convergence may be due to extended periods in the laminar region, the length of which increase with α . The theoretical justification of these results is left for future work.

The estimates found using BIC also show signs of heavy tails and bimodal distributions, particularly for the asymmetric cusp map. The distributions of these estimates are generally more regular, as the higher penalty term in the criterion results in models with fewer parameters fitted, reducing the variability. As with the estimates using AIC, the spread of the estimates for the polynomial map increases with α , again suggesting that the difficulties may be caused by the extended periods in the laminar region. It is unclear why the asymmetric cusp map has such an effect

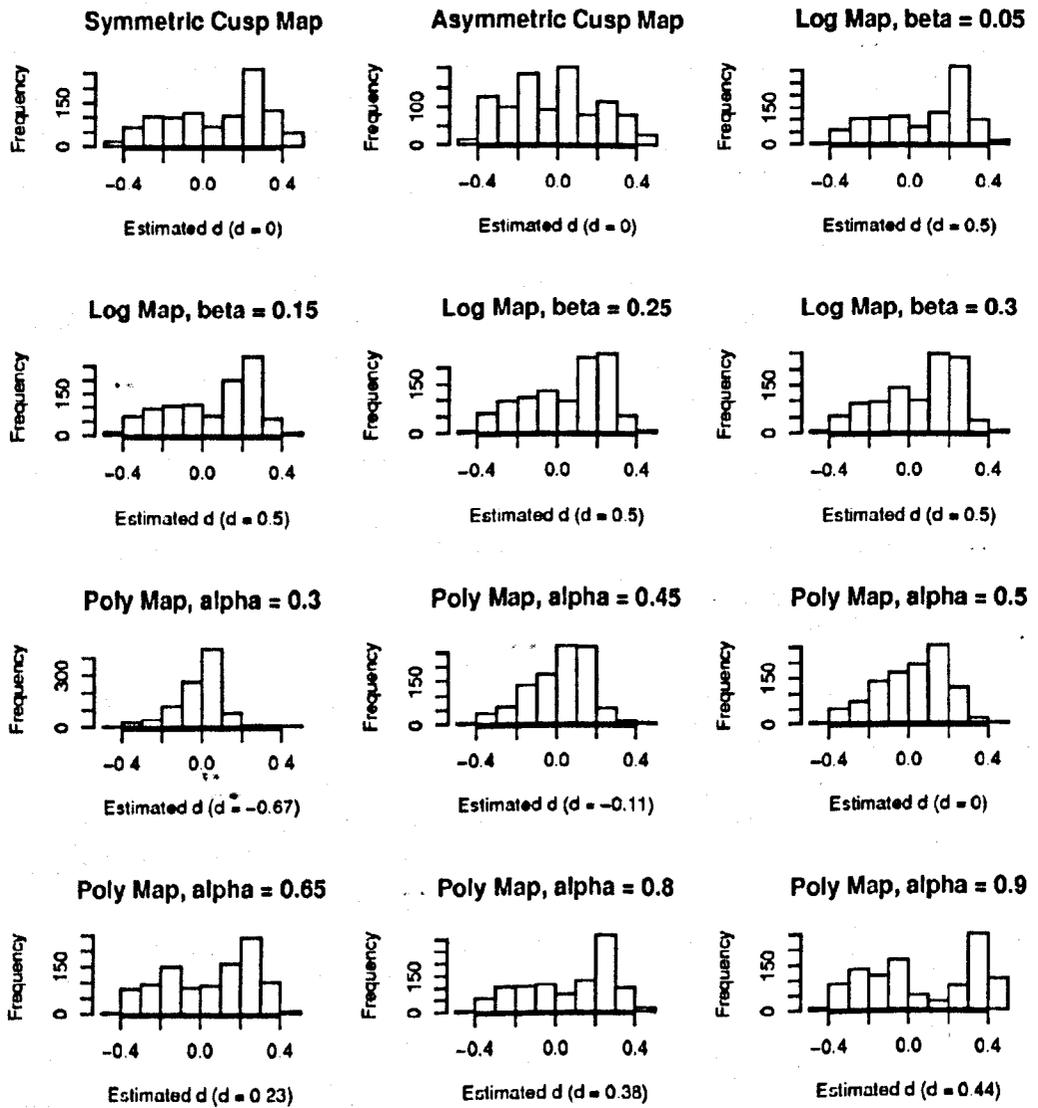


Fig. 4.1: The histograms of the estimates of d for the FARIMA method using AIC

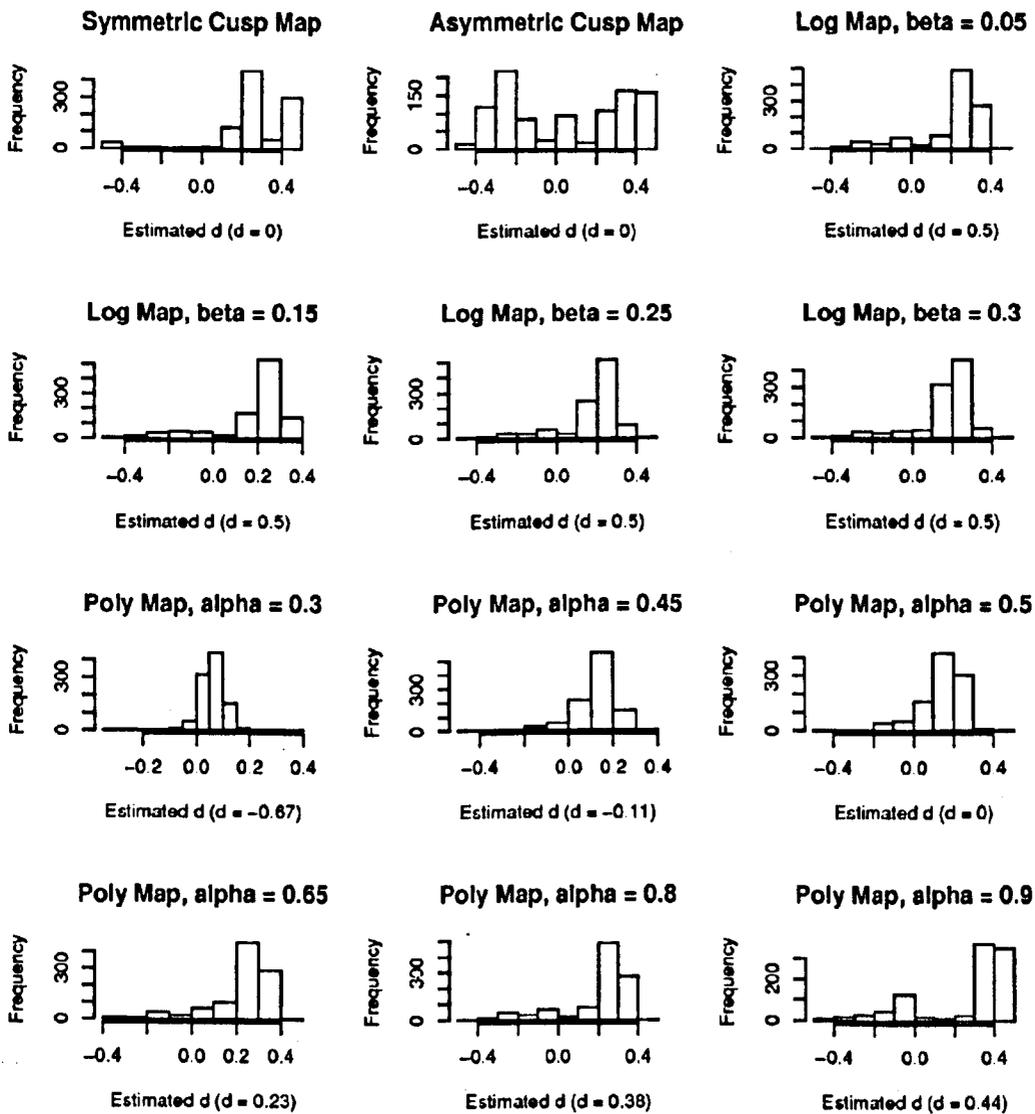


Fig. 4.2: The histograms of the estimates of d for the FARIMA method using BIC

on the estimates, though it seems reasonable to hypothesise that it is related to this map having a second neutral fixed point which is not present in the other maps.

The mean and \sqrt{MSE} of all the methods are given in table 4.1 and the standard deviations are given in table 4.2. The asymptotic values of d shown are based on the asymptotic results for each of the maps, given in section 3.2. The theoretical asymptotic standard deviations are found from the theoretical work presented in Chapter 2. Since several of the methods have theoretical standard deviations which depend on the fitted model, the theoretical standard deviations are computed for each simulation and the intervals given in table 4.2 are the 90% coverage intervals of these theoretical standard deviations, with the upper and lower 5% of theoretical standard deviations removed. It must be noted that many of the assumptions required in deriving these standard deviations do not hold, and these values are included only for reference.

Comparison of the sample standard deviations with the asymptotic standard deviations shows the sample standard deviations tend to be much larger, usually between two to four times the magnitude of the upper limits of the asymptotic deviations, for all methods. For the polynomial map with $\alpha = 0.3$, however, the sample standard deviations from the Local Whittle and GPII methods are reasonably close to the asymptotic results, and for all methods this map produces the estimates with the smallest standard deviations. Since the polynomial map with $\alpha = 0.3$ has orbits which spend the least amount of time, of all the maps considered, in the laminar regions, it gives further evidence to suggest the length of stretches in the laminar region of orbits effects the variability of the estimates produced.

The mean for the FARIMA method using AIC can be seen to be close to zero regardless of the map and as such the \sqrt{MSE} is smaller for maps with d close to zero. This is again due to the general lack of convergence of these estimates, which is reflected in the standard deviations being generally larger for this method than any other, the sole exception being for the asymmetric cusp map.

As expected by the asymptotic results, the mean of \hat{d} using BIC increases with α for the polynomial map. However, as also noted by Bhansali and Holland [2008b], the empirical evidence shows no evidence of intermediate memory for $\alpha < 0.5$. This can also be seen to be true for the other methods of estimation. As expected from figure 4.2, the standard deviations also increase with α .

Also in agreement with Bhansali and Holland [2008b], for the logarithmic map, although the asymptotic value of d is constant, the mean of the estimates can be seen to decrease with β , suggesting the slowly varying function in the correlation decay effects the empirical results. This pattern appears for all estimation methods, although less so for the FARIMA method using AIC. Similarly, the mean estimates for the polynomial map with $\alpha = 0.5$ and the cusp maps could also be effected by the presence of unbounded slowly varying functions in the correlation decays. This is explored further in chapter 5. Although the rate of decay of the correlations for the cusp maps is asymptotically equivalent to the polynomial map with $\alpha =$

Map	Asymptotic d	Mean and \sqrt{MSE} for the Estimated d						
		Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	-0.67	0.017 (0.69)	0.015 (0.69)	-0.012 (0.67)	0.059 (0.73)	0.030 (0.70)	0.064 (0.73)	0.069 (0.74)
Poly (0.45)	-0.11	0.087 (0.21)	0.082 (0.22)	0.016 (0.20)	0.12 (0.25)	0.10 (0.23)	0.15 (0.27)	0.15 (0.26)
Poly (0.5)	0	0.12 (0.14)	0.10 (0.15)	0.018 (0.17)	0.14 (0.18)	0.12 (0.16)	0.18 (0.19)	0.18 (0.19)
Poly (0.65)	0.23	0.22 (0.092)	0.21 (0.12)	0.048 (0.29)	0.22 (0.14)	0.21 (0.15)	0.28 (0.094)	0.28 (0.078)
Poly (0.8)	0.38	0.28 (0.17)	0.28 (0.19)	0.076 (0.37)	0.21 (0.24)	0.19 (0.29)	0.26 (0.18)	0.29 (0.13)
Poly (0.9)	0.44	0.41 (0.11)	0.42 (0.13)	0.072 (0.46)	0.27 (0.28)	0.25 (0.32)	0.33 (0.23)	0.43 (0.089)
Symmetric	0	0.28 (0.29)	0.24 (0.27)	0.072 (0.25)	0.27 (0.32)	0.19 (0.33)	0.32 (0.43)	0.38 (0.38)
Asymmetric	0	0.31 (0.32)	0.26 (0.28)	-0.023 (0.23)	0.044 (0.31)	0.037 (0.31)	-0.069 (0.32)	0.38 (0.39)
Log (0.05)	0.5	0.29 (0.26)	0.29 (0.27)	0.076 (0.48)	0.21 (0.33)	0.19 (0.38)	0.26 (0.28)	0.30 (0.23)
Log (0.15)	0.5	0.26 (0.29)	0.26 (0.30)	0.057 (0.49)	0.18 (0.36)	0.17 (0.39)	0.24 (0.29)	0.28 (0.25)
Log (0.25)	0.5	0.23 (0.31)	0.24 (0.32)	0.052 (0.49)	0.18 (0.35)	0.17 (0.38)	0.22 (0.31)	0.25 (0.27)
Log (0.3)	0.5	0.22 (0.33)	0.22 (0.34)	0.054 (0.49)	0.17 (0.36)	0.16 (0.38)	0.21 (0.31)	0.24 (0.29)

Tab. 4.1: The values in brackets are the \sqrt{MSE} . All values are given correct to 2 significant figures.

Estimated Standard Deviations of \hat{d}							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.053	0.074	0.118	0.050	0.056	0.041	0.044
Poly (0.45)	0.073	0.097	0.151	0.091	0.083	0.055	0.058
Poly (0.5)	0.082	0.106	0.169	0.110	0.111	0.063	0.061
Poly (0.65)	0.092	0.122	0.224	0.144	0.153	0.079	0.064
Poly (0.8)	0.138	0.167	0.219	0.161	0.211	0.137	0.102
Poly (0.9)	0.105	0.124	0.272	0.227	0.260	0.203	0.088
Symmetric	0.090	0.119	0.239	0.187	0.270	0.284	0.076
Asymmetric	0.083	0.111	0.226	0.302	0.309	0.309	0.088
Log (0.05)	0.138	0.167	0.219	0.161	0.211	0.137	0.102
Log (0.15)	0.141	0.174	0.213	0.166	0.210	0.129	0.098
Log (0.25)	0.147	0.179	0.202	0.141	0.190	0.123	0.099
Log (0.3)	0.145	0.180	0.192	0.137	0.181	0.118	0.099

Asymptotic Standard Deviations from Fitted Models							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.040	0.064	(0.023,0.067)	(0.021,0.043)	(0.022,0.030)	(0.021,0.026)	(0.018,0.027)
Poly (0.45)	0.040	0.064	(0.024,0.064)	(0.019,0.038)	(0.022,0.032)	(0.019,0.028)	(0.018,0.027)
Poly (0.5)	0.040	0.064	(0.024,0.057)	(0.019,0.039)	(0.022,0.033)	(0.018,0.028)	(0.015,0.027)
Poly (0.65)	0.040	0.064	(0.022,0.060)	(0.017,0.043)	(0.020,0.032)	(0.017,0.026)	(0.015,0.027)
Poly (0.8)	0.040	0.064	(0.020,0.077)	(0.016,0.036)	(0.016,0.028)	(0.016,0.022)	(0.015,0.027)
Poly (0.9)	0.040	0.064	(0.022,0.057)	(0.016,0.038)	(0.019,0.028)	(0.015,0.024)	(0.015,0.044)
Symmetric	0.040	0.064	(0.023,0.084)	(0.012,0.037)	(0.012,0.036)	(0.008,0.021)	(0.010,0.029)
Asymmetric	0.040	0.064	(0.035,0.061)	(0.017,0.044)	(0.023,0.042)	(0.017,0.032)	(0.021,0.036)
Log (0.05)	0.040	0.064	(0.019,0.061)	(0.016,0.040)	(0.016,0.029)	(0.015,0.023)	(0.015,0.027)
Log (0.15)	0.040	0.064	(0.020,0.072)	(0.016,0.041)	(0.016,0.029)	(0.015,0.022)	(0.015,0.027)
Log (0.25)	0.040	0.064	(0.019,0.049)	(0.016,0.036)	(0.016,0.029)	(0.015,0.022)	(0.015,0.027)
Log (0.3)	0.040	0.064	(0.019,0.055)	(0.016,0.034)	(0.016,0.029)	(0.015,0.022)	(0.015,0.027)

Tab. 4.2: The value in brackets are the 90% coverage intervals of the asymptotic standard deviations based on the fitted models, disregarding the lower and upper 5%. For the Local Whittle and GPH methods, the asymptotic standard deviation depends only on m , which was fixed, and is thus quoted as a single number.

0.5, the results show a stronger positive bias for the cusp map, suggesting a slower convergence to the asymptotic results.

Since the polynomial map with $\alpha \neq 0$ have spectral densities of a more standard form, it may reasonably be expected that the estimates from these maps outperform the others. For $\alpha > 0.5$, the values of the sample means and \sqrt{MSE} s reported in table 4.1 shows the empirical evidence appears to support this claim. The means for the local Whittle, GPII and FExp methods for the polynomial map with $\alpha = 0.65$ and 0.9 in particular seem very close to the asymptotic results. The means for the polynomial map with $\alpha = 0.8$ seem negatively biased for all methods, although it is not clear why this is the case. The FAR method, using both AIC and BIC, seems to perform well in terms of mean for the polynomial map with $\alpha = 0.65$, although higher values of α leads to larger negative bias.

The means for all of the maps for the FARIMA method using BIC seem bounded away from $d = 0$ and $d = 0.5$, producing positive bias for the cusp maps and polynomial map with $\alpha \leq 0.5$ and negative bias for $\alpha \geq 0.65$ and the logarithmic maps.

The histograms of the estimates of d for the FAR method using AIC and the FEXP are given in figures 4.3 and 4.4 respectively. The histograms for the FAR method using BIC are similar to those in figure 4.3 and thus omitted. As semi-parametric methods, these estimates are expected to be more robust than the FARIMA method to misspecification.

The estimates found for the FAR method using AIC still show signs of heavy tails and bimodal distributions and are very similar in appearance to those for the FARIMA method using BIC. Note in particular that the same difficulties occur for the asymmetric cusp map and that as with the FARIMA method, the distributions become more irregular for the polynomial map as α increases. The FAR method using BIC shows less signs of heavy tails and thus comparing like for like the FAR method does reduce these difficulties compared to the FARIMA method, as expected.

Although some skewness can be seen, the distributions for the FEXP estimates are more regular for all of the maps, showing no signs of bimodal distributions. They are not, however, normally distributed. Indeed, the Kolmogorov-Smirnov test for Normality shows that the distributions of the estimates for all methods over all maps are significantly non-Gaussian at the 0.1% significance level.

As with the FARIMA method, table 4.1 shows the mean and standard deviation of \hat{d} to increase with α for the polynomial map and the mean to decrease with β for the logarithmic map. There are also signs of the standard deviation also decreasing with β .

To investigate more closely the difference in behaviours of the FAR and FExp methods, a single orbit of length 10000 from the polynomial map with $\alpha = 0.9$ is taken. For this orbit, the FExp estimate of d is 0.46 with s fitted as 5, whilst the FAR estimate using AIC is -0.48, with p fitted as 5. The raw periodogram along with the theoretical spectral densities for the fitted FAR model using AIC and the fitted FExp model are shown in figure 4.5. It can be seen that despite giving an estimate

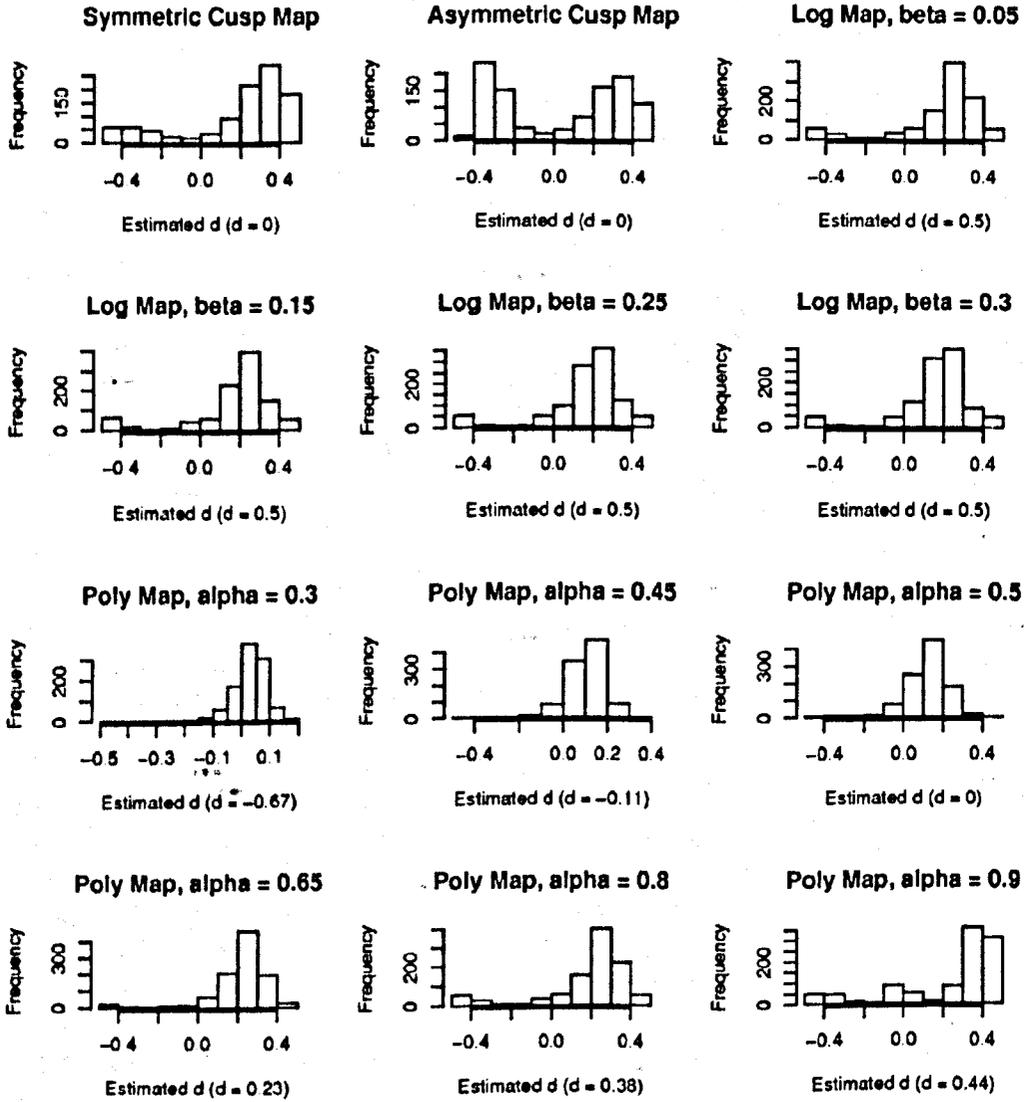


Fig. 4.3: The histograms of the estimates of d for the FAR method using AIC

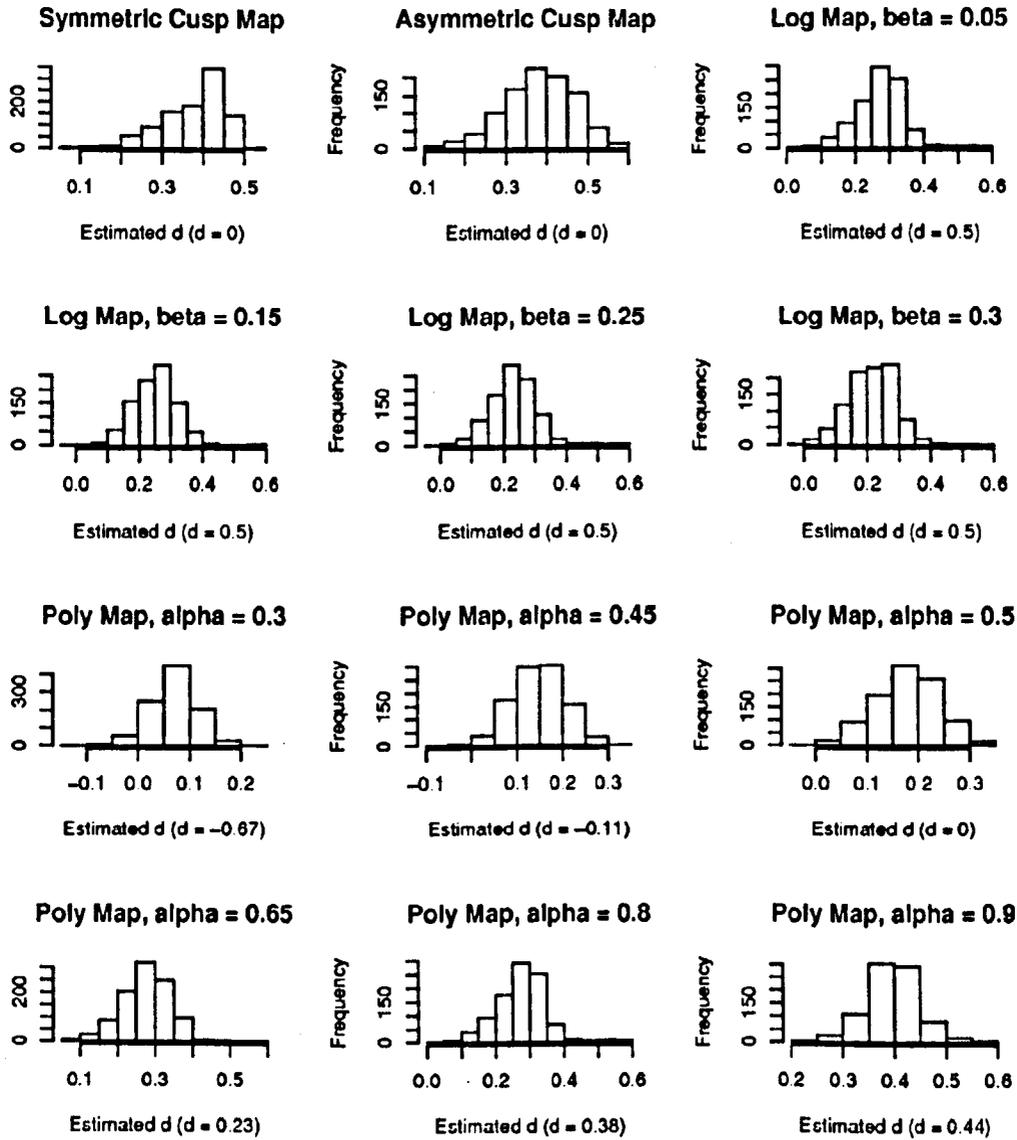


Fig. 4.4: The histograms of the estimates of d for the FEXP method

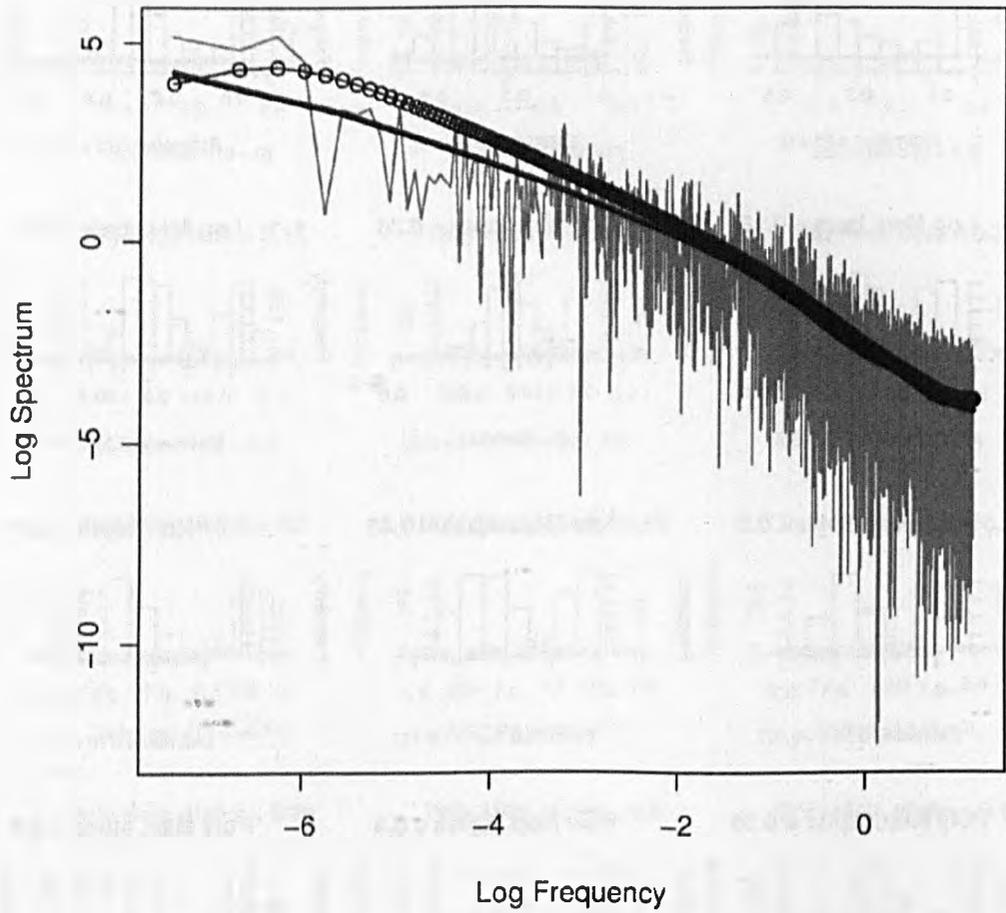


Fig. 4.5: Log-Log plots of the raw spectral density from an orbit of the polynomial map with $\alpha = 0.9$. The bold line is the theoretical spectrum from the fitted FExp model and the line with circles is from the fitted FAR model using AIC.

of d as -0.48 and thus suggesting strong intermediate memory, the FAR spectral estimate still appears close to the raw periodogram, although a small decrease can be seen near $\lambda = 0$. Taking the

$$\sqrt{MSE} = \sqrt{\frac{1}{5000} \sum_{j=1}^{5000} (\hat{f}(\lambda_j) - I(\lambda_j))^2}$$

where $I(\lambda_j)$ is the raw periodogram gives the \sqrt{MSE} for the FAR spectral estimate as 4.06 whilst the \sqrt{MSE} for the FExp estimate is 5.51 , thus the FAR spectral estimate indeed seems to fit the periodogram better. This suggests the short memory component of the FAR model fitted is attempting to model the long memory behaviour, altering the estimated value of d . A similar comment was made by Bhansali and Kokoska [2001a], which suggested extra parameters caused by overfitting attempted to model the long memory, altering the estimate of d .

The histograms for the non-parametric methods are close to normal in appearance and thus omitted. In replacement, normal Q-Q plots are provided for the local Whittle method in figure 4.6. The Q-Q plots for the GPII method are similar in appearance and thus not shown.

It can be seen from figure 4.6 that the distributions of the estimates from the local Whittle method and, although not shown, the GPII method are close to normal other than near the tails. It can also be seen that the shape of the distributions does not greatly vary between maps, nor are they strongly dependent on α or β . This gives further evidence to the greater robustness of non-parametric methods in comparison to parametric and semi-parametric methods.

However, from table 4.2 it can be seen that as with the other methods, the standard deviations of the non-parametric methods increase with α for the polynomial map. It can also be seen that for the majority of maps, the standard deviations for the non-parametric method are greater than those for the FAR method using BIC and the FEXP method, whilst table 4.1 shows the mean biases are often of similar magnitude.

4.2 Further Simulation Studies

Since the results of the simulation study carried out in section 4.1 generally shows a lack of convergence of the orbits from the chaotic intermittency maps to their asymptotic properties values of d , two further simulation studies are carried out here to provide further empirical evidence that these asymptotic theories hold.

These new simulation studies follow the same general structure of the previous simulation study, described in section 4.1.1. The deviations from this outline for each are explained.

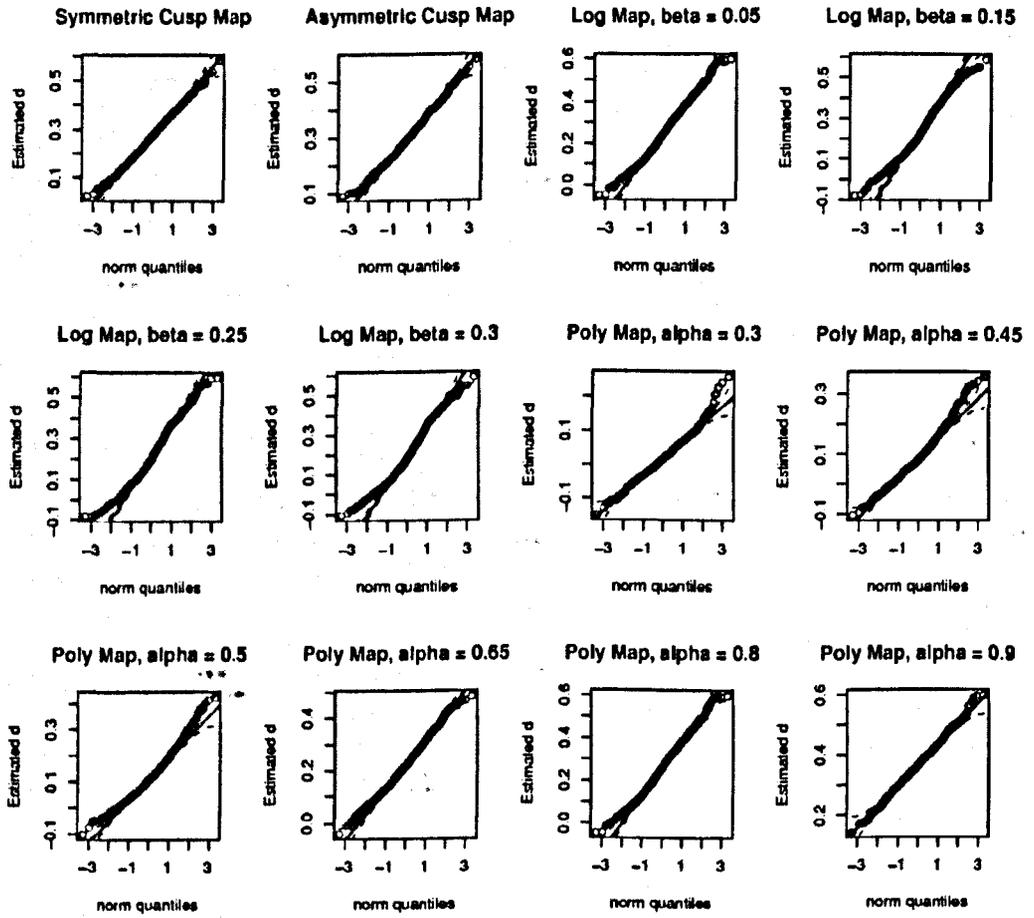


Fig. 4.6: The normal Q-Q plots of the estimates of d for the local Whittle method

4.2.1 Extended Length Simulation Study

The extended length simulation study attempts to discover if the estimated values of d tend to converge to the asymptotic results as the series length increases. As such, the initial values for each simulation were generated as described in section 4.1.1 and the first $M = 10^7 - 10^4$ iterations were once again disregarded. The length of the orbit retained for study, T , was then allowed to increase up to $T = 10^6$ in 10 steps of 10^5 . Due to the large computational time required, only 100 repetitions of each map were generated and only the local Whittle and GPH methods of estimation were applied, as these require the least computational time. For both of these methods m was taken as $T^{1/2}$. The results for both the local Whittle and GPH methods in terms of mean, standard deviation and distribution are very similar and thus only the results for the GPH method are reported.

Figure 4.7 shows the development of the normal Q-Q plots for the GPH estimates from the logarithmic map with $\beta = 0.05$ as T increases. It can be seen that as T increases the distributions become closer to normality and there is no significant evidence to reject normality of the distribution at the 0.1% significance level for $T = 1,000,000$. This is also true for the estimates from the other maps, the plots for which are omitted.

Figure 4.8 shows the mean values of the GPH estimates for the polynomial maps as T increases. It can be seen that for the polynomial map with $\alpha < 0.5$ the estimates do not appear to be decreasing towards $1 - 1/2\alpha$, but instead seem to remain near zero as T increases. Hence, even at $T = 1,000,000$ no evidence of intermediate memory can be seen.

For the polynomial map with $\alpha = 0.5$ the GPH estimates do appear to decrease slightly towards zero, but at a very slow rate. Since the theoretical spectral density for this map still has a singularity at zero this may be expected. Neither the local Whittle or the GPH methods are applicable when such boundary behaviour is present. Chapter 5 investigates the bias caused by the boundary behaviour further.

For the polynomial map with $\alpha > 0.5$ the GPH estimates seem to tend towards the asymptotic values of d . Although the mean estimates are still below the values given by the asymptotic theory, this value is consistently within 1 standard deviation of the mean. This suggests that the asymptotic results of the polynomial map with $\alpha > 0.5$ seem to hold and also that the GPH method can provide unbiased estimates of d for these non-linear, non-Gaussian series. The same is true for the local Whittle estimates.

Figure 4.9 shows the mean values of the GPH estimates for the symmetric and asymmetric cusp maps and the logarithmic maps as T increases. Similar to the polynomial map with $\alpha = 0.5$, the GPH estimates for the asymmetric and symmetric cusp maps appear to possibly be decreasing towards the asymptotic value of $d = 0$, but at a very slow rate. As with the polynomial map with $\alpha = 0.5$, the cusp maps possess lower boundary long memory behaviour, or 'weak' long memory, which the GPH and local Whittle methods are not designed for, explored further in chapter 5.

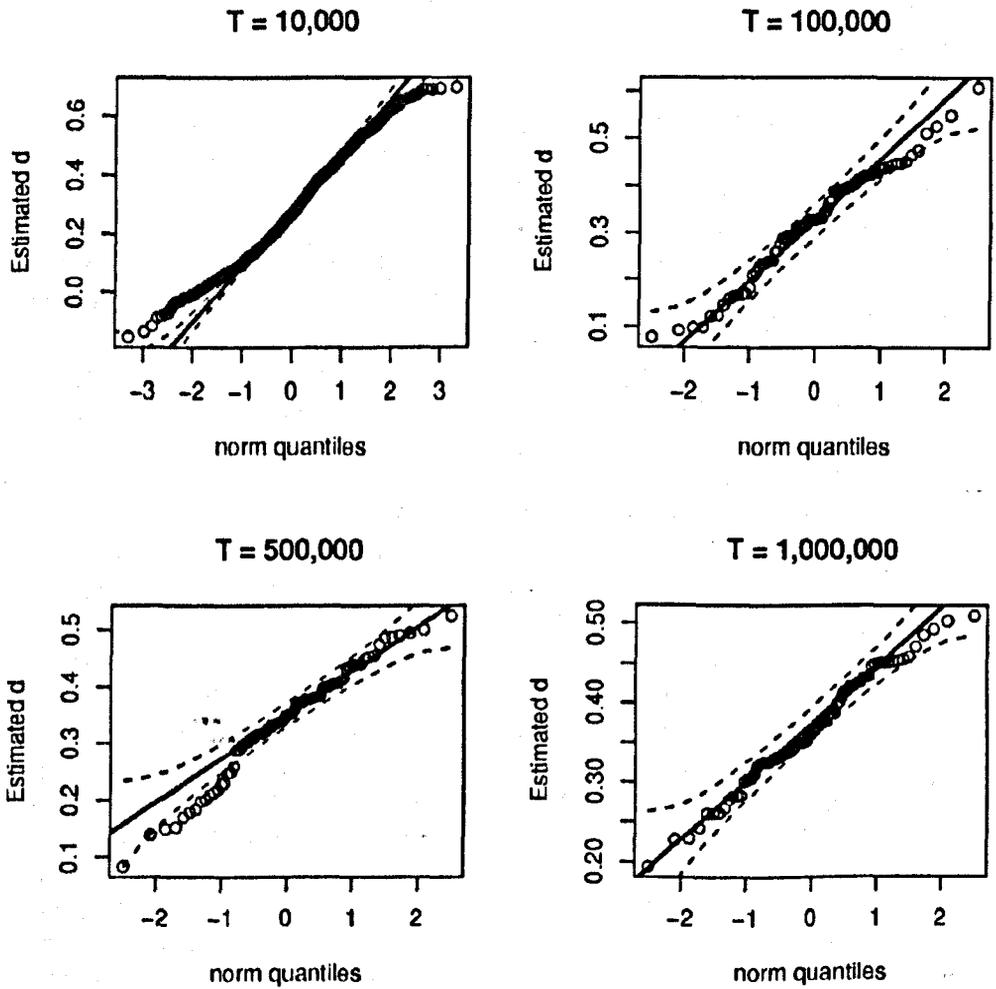


Fig. 4.7: Normal Q-Q plots for the GPII estimates from the logarithmic map with $\beta = 0.05$ as T increases.

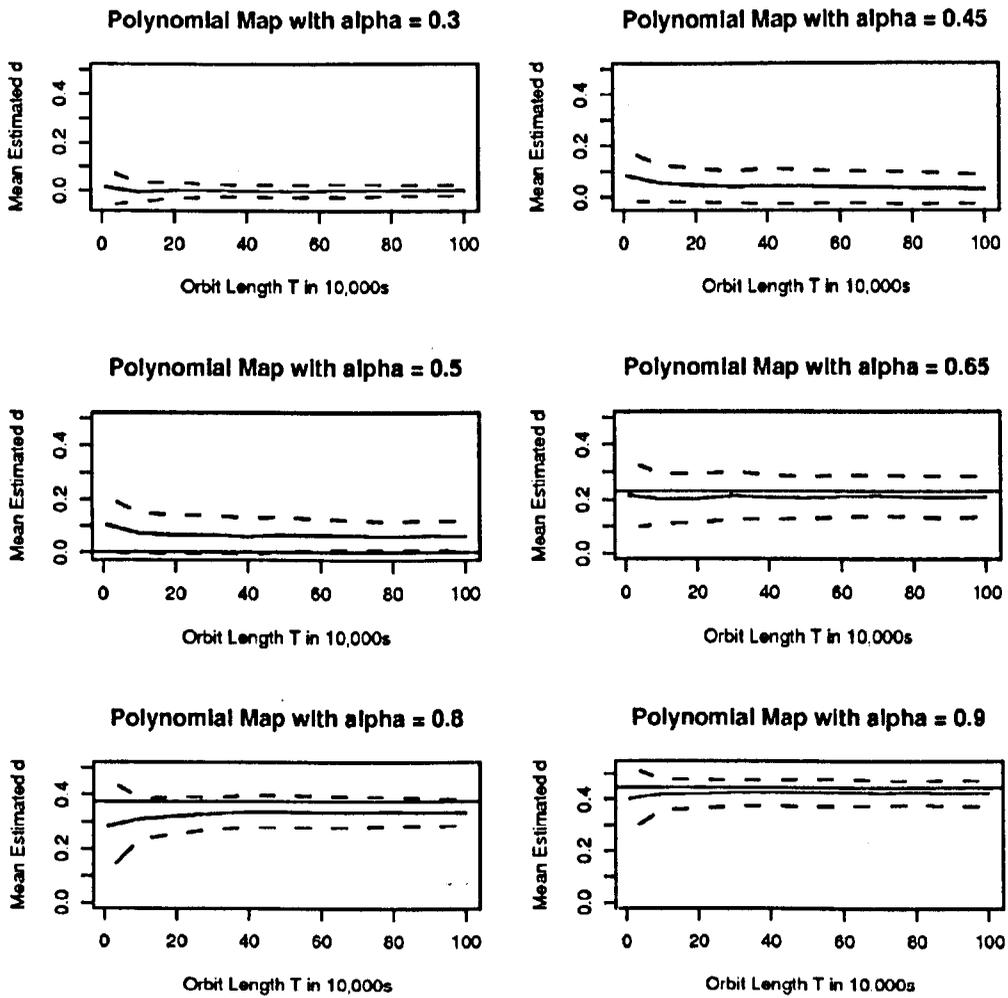


Fig. 4.8: The mean values of the GPII estimates for the Polynomial maps as T increases. The dotted lines represent the mean $\pm \hat{\sigma}$. The horizontal line represents the asymptotic value of d .

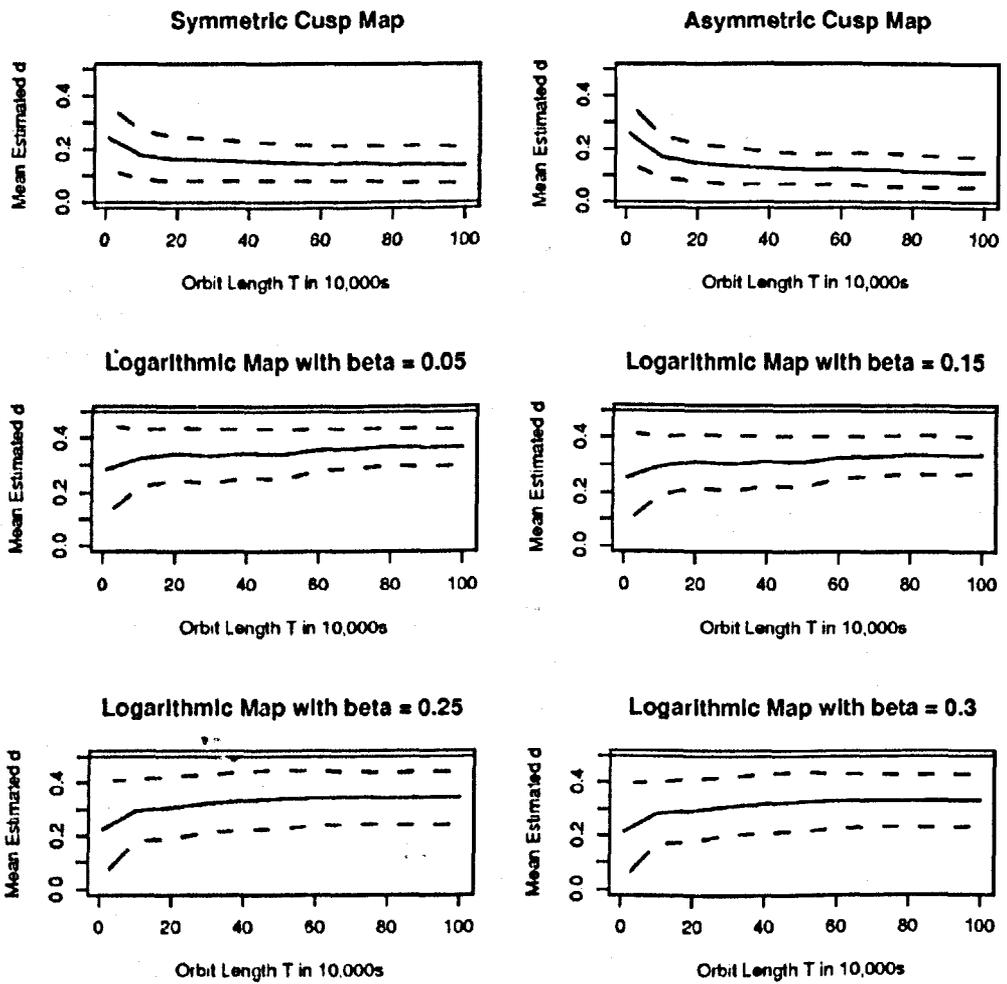


Fig. 4.9: The mean values of the GPII estimates for the Symmetric and Asymmetric cusp maps and the Logarithmic maps as T increases. The dotted lines represent the mean $\pm \sigma$. The horizontal line represents the asymptotic value of d .

The GPH estimates for the logarithmic maps appear to increase with T , reducing the magnitude of the bias, but again at a slow rate. The results also still show dependency on β . The logarithmic maps display behaviour at the upper boundary of long memory and the amount of bias accounted for by this behaviour will be investigated further in Chapter 5.

Table 4.3 shows the sample standard deviations for the GPH estimates as the series length increases. The asymptotic values quoted are $\pi/\sqrt{24m}$, where m has been chosen as $m = T^{1/2}$, see Hurvich et al [1998]. For the polynomial map with $\alpha = 0.3$ the sample standard deviations are very close to those suggested by the asymptotic theory and indeed seem to get closer as T increases. In general, the sample standard deviations of the local Whittle estimates from the chaotic intermittency maps are very similar to those reported in table 4.3 for the GPH estimates with the ratio of the two seeming to tend towards 1 and are thus omitted. The only exception to this is for the polynomial map with $\alpha = 0.3$, for which the ratio is close to the asymptotic ratio of $\pi/\sqrt{6}$. The results thus seem to show that the asymptotic theory regarding the standard deviations for both the GPH method and the local Whittle method still holds for the polynomial map with $\alpha = 0.3$. This also suggests that the large bias of the estimates may be due more to a lack of convergence of the map to its theoretical properties than a failure of the estimation methods.

For the other chaotic intermittency maps, including the polynomial maps with $\alpha \geq 0.45$, although the sample standard deviations seem to decrease as T increases, they do so at a slower rate than suggested by the asymptotic theory. For the polynomial map with $\alpha > 0.5$, the decrease in sample standard deviations suggests consistency of the GPH and local Whittle estimates still holds.

The standard deviations for the polynomial maps appear to be smaller than the other maps. The standard deviations for the logarithmic maps are seen to be the largest and appear to increase with β .

4.2.2 Systematic Sampling Simulation Study

Systematic sampling of a series refers to producing a new series by taking every n^{th} observation of the original series. Work done by Chambers [1998] shows this new series will have the same asymptotic d value as the original. Ilwang [2000], however, shows a reduction in the estimated d for a finite sample can be expected.

In the chaotic intermittency maps, the main effect of systematic sampling will be to break up long periods in the laminar region. The overall amount of time spent in the laminar regions should remain approximately unchanged and the asymptotic values of d will be unaffected, but the reduction of long periods of consecutive values within the laminar region may improve the behaviour of the estimates.

The initial values for each simulation were generated as described in section 4.1.1. The first $M = 10^7 - 10^4$ iterations were once again disregarded. The length of the orbit retained for study, T , was then taken as $T = 10^6$. Two new series were then

Map	Length T in 10.000s										
	1	10	20	30	40	50	60	70	80	90	100
Poly (0.3)	0.074	0.039	0.033	0.026	0.024	0.025	0.025	0.027	0.022	0.022	0.021
Poly (0.45)	0.097	0.072	0.066	0.062	0.070	0.067	0.062	0.062	0.064	0.058	0.056
Poly (0.5)	0.106	0.076	0.071	0.070	0.069	0.064	0.059	0.055	0.055	0.056	0.056
Poly (0.65)	0.122	0.091	0.090	0.086	0.082	0.076	0.077	0.074	0.074	0.076	0.074
Poly (0.8)	0.167	0.075	0.069	0.058	0.060	0.059	0.059	0.056	0.053	0.052	0.047
Poly (0.9)	0.124	0.057	0.056	0.050	0.049	0.053	0.051	0.048	0.046	0.049	0.049
Symmetric	0.119	0.091	0.084	0.079	0.073	0.070	0.067	0.066	0.068	0.070	0.068
Asymmetric	0.111	0.079	0.070	0.070	0.060	0.057	0.058	0.063	0.061	0.058	0.057
Log (0.05)	0.167	0.106	0.097	0.100	0.091	0.092	0.077	0.073	0.070	0.071	0.067
Log (0.15)	0.174	0.105	0.107	0.105	0.101	0.102	0.090	0.097	0.084	0.085	0.083
Log (0.25)	0.179	0.115	0.119	0.108	0.109	0.111	0.100	0.096	0.095	0.097	0.096
Log (0.3)	0.180	0.117	0.117	0.109	0.111	0.111	0.103	0.099	0.098	0.100	0.099
Asymptotic	0.064	0.036	0.030	0.027	0.025	0.024	0.023	0.022	0.021	0.021	0.020

Tab. 4.3: The sample standard deviations for the GPH estimates as the series length increases. The asymptotic values are $\pi/\sqrt{24m}$ and $m = T^{1/2}$.

taken from this orbit. The first, s_t^{10} , was created by taking every 10^{th} observation of the original series up to $T = 10^5$. The second, s_t^{100} , was created by taking every 100^{th} observation of the original series up to $T = 10^6$. The lengths of each of s_t^{10} and s_t^{100} are thus 10^4 . For each of the maps, 1000 such s_t^{10} and s_t^{100} series were generated, and the long memory parameter for each was recorded using the methods as described in section 4.1.1.

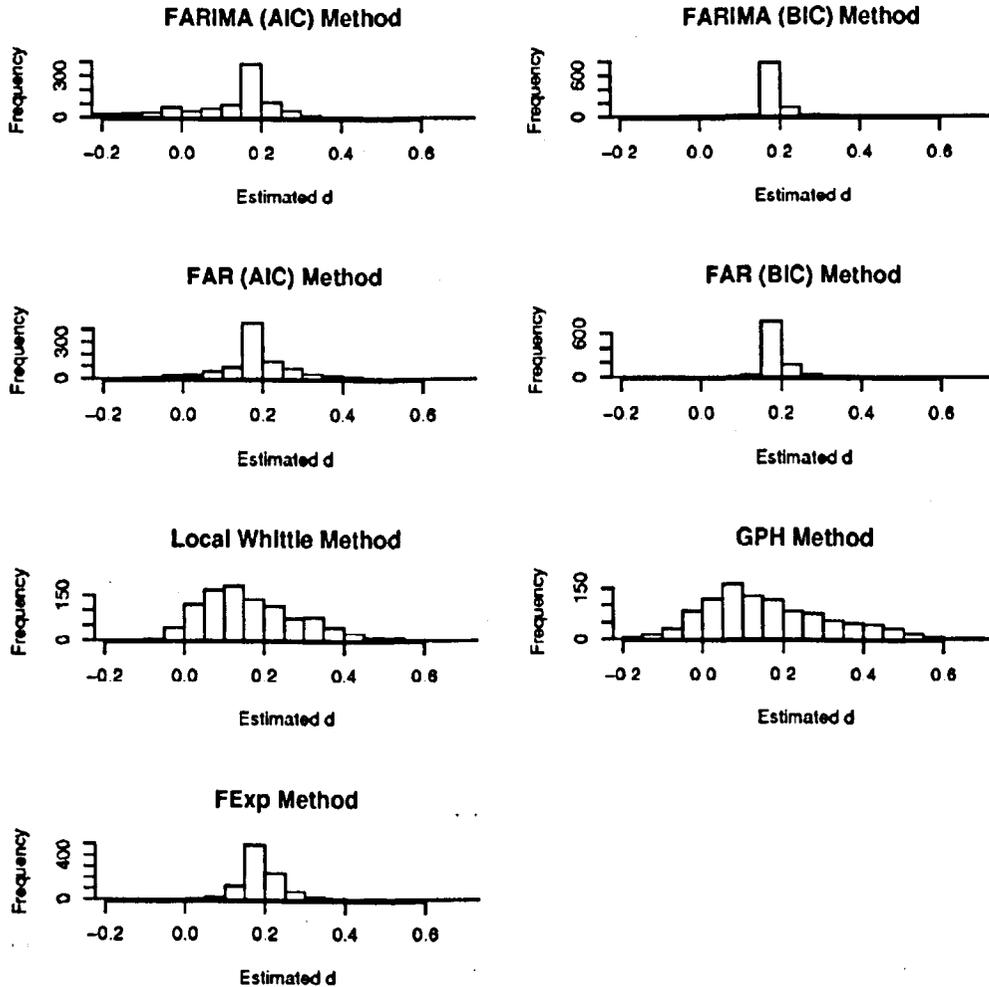


Fig. 4.10: The histograms of the estimates of the long memory parameter for the $s_t^{(10)}$ series from the polynomial map with $\alpha = 0.65$

Figure 4.10 gives the histograms of the estimates of the long memory parameter for the $s_t^{(10)}$ series from the polynomial map with $\alpha = 0.65$. In comparison with the results given in section 4.1, the FAR and FARIMA methods now appear to converge

much more to a single value of d . The distributions of the FAR and FARIMA estimates seem much closer to the distribution of the FExp estimates. The distributions of the local Whittle and GPI estimates appear to have become more asymmetric, with positive skew. It can be seen that all of the distributions are significantly different from normal, particularly in the tail behaviour. The distributions for the other maps are similar in appearance, being significantly different from normal with heavy tails and thus the figures are omitted.

The sample means of the estimates of d for the $s_t^{(10)}$ and $s_t^{(100)}$ series are given in table 4.4. For the polynomial map with $\alpha < 0.5$, the sample means from both the $s_t^{(10)}$ and $s_t^{(100)}$ series are close to zero for all methods. This is similar to the results given in section 4.1 and still no evidence of intermediate memory is seen. For the polynomial map with $\alpha = 0.5$, the sample means are also close to zero. This reduction in bias is expected, however, by the work of Hwang [2000], which shows systematic sampling reduces the estimated d for a finite sample. Comparison of the sample means for all the maps for the $s_t^{(10)}$ series with the sample means from the $s_t^{(100)}$ series shows a reduction in the sample means of the estimates in almost every case.

For the polynomial map with $\alpha > 0.5$, the sample means still appear to increase with α , although a negative bias can be seen throughout. Note, that for the polynomial map with $\alpha = 0.8$, the sample means have actually increased for the $s_t^{(10)}$ series compared to those reported in section 4.1, causing a reduction in bias. The sample means for the estimates using the FARIMA (AIC) method have also increased for the polynomial maps with $\alpha = 0.65$ and $\alpha = 0.9$ showing better convergence and reduction in bias. Such results are not anticipated by the general theory of the effect of systematic sampling.

For both the cusp maps, the sample means of the local Whittle, GPI and FExp estimates have decreased significantly compared to those reported in section 4.1. The sample means for the $s_t^{(100)}$ series are close to zero for all of the estimation techniques.

The \sqrt{MSE} for the estimates for the $s_t^{(10)}$ and $s_t^{(100)}$ series are given in table 4.5. It can be seen for the polynomial map with $\alpha \leq 0.5$ the \sqrt{MSE} has been reduced for every estimation method, except the FARIMA (AIC) method for the polynomial map with $\alpha = 0.3$ for which it has remained almost the same. This result is not surprising however, since the original estimates showed positive bias and systematic sampling has been shown to reduce estimates. More significantly, systematic sampling has reduced the \sqrt{MSE} for all of the estimation methods for the polynomial map with $\alpha = 0.8$ for the $s_t^{(10)}$ series and the majority of estimates from the FARIMA and FAR methods using both AIC and BIC and the FExp method have also shown reduction in \sqrt{MSE} for the polynomial maps with $\alpha = 0.65$ and

Map	Asymptotic d	Mean for the Estimated d for $s_t^{(10)}$ series.						
		Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	-0.67	-0.008	0.003	-0.001	0.023	0.004	0.023	0.008
Poly (0.45)	-0.11	0.012	0.016	0.015	0.068	0.038	0.070	0.053
Poly (0.5)	0	0.040	0.042	0.038	0.097	0.068	0.098	0.084
Poly (0.65)	0.23	0.167	0.166	0.109*	0.181	0.162	0.184	0.196
Poly (0.8)	0.38	0.297*	0.302*	0.151*	0.241*	0.236*	0.254	0.300*
Poly (0.9)	0.44	0.357	0.369	0.179*	0.272	0.249	0.286	0.360
Symmetric	0	0.114	0.103	0.075*	0.191	0.136	0.209	0.189
Asymmetric	0	0.118	0.106	0.043*	0.146*	0.125*	0.219*	0.204
Log (0.05)	0.5	0.283	0.289	0.097*	0.197	0.180	0.196	0.275
Log (0.15)	0.5	0.254	0.261	0.072*	0.183	0.166	0.180	0.253
Log (0.25)	0.5	0.223	0.230	0.065*	0.163	0.164	0.165	0.227
Log (0.3)	0.5	0.217	0.224	0.064*	0.149	0.135	0.157	0.220
Map	Asymptotic	Mean for the Estimated d for $s_t^{(100)}$ series.						
Poly (0.3)	-0.67	-0.007	0.002	-0.008	-0.001	-0.003	-0.001	-0.001
Poly (0.45)	-0.11	-0.006	0.004	-0.003	0.005	0.001	0.006	0.004
Poly (0.5)	0	-0.003	0.008	0.000	0.013	0.008	0.012	0.012
Poly (0.65)	0.23	0.071	0.076	0.039	0.070	0.078	0.071	0.088
Poly (0.8)	0.38	0.246	0.255	0.097*	0.184	0.179	0.183	0.233
Poly (0.9)	0.44	0.325	0.340	0.134*	0.234	0.231	0.248	0.310
Symmetric	0	0.022	0.026	0.019	0.053	0.043	0.054	0.054
Asymmetric	0	0.024	0.029	0.026*	0.083*	0.053*	0.082*	0.072
Log (0.05)	0.5	0.248	0.259	0.054	0.152	0.158	0.163	0.237
Log (0.15)	0.5	0.213	0.224	0.033	0.131	0.148	0.149	0.211
Log (0.25)	0.5	0.181	0.189	0.030	0.107	0.124	0.121	0.185
Log (0.3)	0.5	0.167	0.174	0.029	0.097	0.119	0.110	0.168

Tab. 4.4: The sample means of the estimates of d for the $s_t^{(10)}$ and $s_t^{(100)}$ series. The starred values are those which have increased after systematic sampling has been applied.

\sqrt{MSE} of \hat{d} for $s_t^{(10)}$ series.							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.661*	0.674*	0.669	0.690*	0.672*	0.689*	0.675*
Poly (0.45)	0.140*	0.154*	0.158*	0.182*	0.160*	0.183*	0.169*
Poly (0.5)	0.096*	0.119*	0.109*	0.107*	0.100*	0.103*	0.097*
Poly (0.65)	0.136	0.169	0.187*	0.073*	0.155	0.084*	0.076*
Poly (0.8)	0.128*	0.154*	0.298*	0.166*	0.250*	0.168*	0.100*
Poly (0.9)	0.120	0.135	0.346*	0.218*	0.324	0.216*	0.113
Symmetric	0.152*	0.161*	0.176*	0.202*	0.180*	0.211*	0.198*
Asymmetric	0.156*	0.164*	0.183*	0.207*	0.183*	0.231*	0.218*
Log (0.05)	0.266	0.284	0.442*	0.323	0.398	0.347	0.252
Log (0.15)	0.295	0.310	0.464*	0.339	0.406	0.360	0.274
Log (0.25)	0.324	0.337	0.467*	0.355	0.394	0.367	0.299
Log (0.3)	0.332	0.344	0.466*	0.370	0.427	0.376	0.307
\sqrt{MSE} of \hat{d} for $s_t^{(100)}$ series.							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.661*	0.673*	0.661*	0.666*	0.664*	0.666*	0.666*
Poly (0.45)	0.118*	0.139*	0.121*	0.117*	0.116*	0.118*	0.118*
Poly (0.5)	0.052*	0.076*	0.063*	0.020*	0.031*	0.018*	0.027*
Poly (0.65)	0.194	0.207	0.219*	0.170	0.181	0.168	0.159
Poly (0.8)	0.177	0.196	0.326*	0.214	0.281	0.226	0.170
Poly (0.9)	0.155	0.166	0.364*	0.243	0.320	0.236	0.164
Symmetric	0.087*	0.103*	0.090*	0.061*	0.076*	0.065*	0.075*
Asymmetric	0.083*	0.098*	0.100*	0.089*	0.088*	0.090*	0.088*
Log (0.05)	0.301	0.316	0.480	0.375	0.410	0.376	0.293
Log (0.15)	0.334	0.344	0.495	0.392	0.411	0.380	0.319
Log (0.25)	0.365	0.375	0.494	0.410	0.425	0.405	0.344
Log (0.3)	0.375	0.385	0.496	0.416	0.423	0.413	0.357

Tab. 4.5: The \sqrt{MSE} for the estimates for the $s_t^{(10)}$ and $s_t^{(100)}$ series. The starred values represent a \sqrt{MSE} which has decreased after systematic sampling.

$\alpha = 0.9$. The results from section 4.1 showed a negative bias for these maps and systematic sampling would be expected to increase this bias.

For the cusp maps, Table 4.5 shows the \sqrt{MSE} s have all been reduced by systematic sampling for the both the cusp maps and table 4.6 shows a reduction in variance for FARIMA and FAR methods using both AIC and BIC and the FExp method. This further suggests the benefits of systematic sampling to reduce the length of consecutive values in the laminar region.

The sample standard deviations for the estimates for the $s_t^{(10)}$ and $s_t^{(100)}$ series are given in table 4.6. The magnitudes of the sample standard deviations are approximately the same for both systematic sampling approaches. For the polynomial maps it can be seen that systematic sampling gives a reduction in sample standard deviations for almost every case. This reinforces the idea that too many consecutive values in the laminar region leads to the large variances shown in section 4.1 and that systematic sampling can reduce this variability.

Apart from slight increases for the FARIMA method using AIC for the $s_t^{(10)}$ series, the sample means of the estimates for the logarithmic maps given in table 4.4 show a decrease compared to those given in section 4.1, which increases the negative bias. Tables 4.5 and 4.6 also show increases in the \sqrt{MSE} and sample standard deviations for the local Whittle, GPII, FAR and FExp estimates. For the FARIMA method using both AIC and BIC reductions in the \sqrt{MSE} and sample standard deviations can be seen, implying systematic sampling can still improve these estimates. However, in general systematic sampling does not appear to improve the estimates for the logarithmic maps, suggesting large sequences of consecutive values within the laminar region are not the main cause of difficulty in estimating d for these maps.

4.3 Conclusions

As previously mentioned by Bhansali and Holland [2008b], the finite sample properties of the chaotic intermittency maps only partly agree with the asymptotic theory. In addition, the methods of estimating d applied are not suitable for the boundary cases of long memory displayed by several of the maps.

For the polynomial map with $\alpha < 0.5$, no empirical evidence of the theoretical intermediate memory was found here or in the simulation studies of Bhansali, Holland and Kokoska [2006] and Bhansali and Holland [2008b]. The extended length simulation study, section 4.2.1, showed that if the orbits do tend towards intermediate memory they appear to do so at a very slow rate, with no signs at a series length of 10^6 . It seems plausible that although the asymptotic theory has shown regions of the polynomial maps with $\alpha < 0.5$ which possess intermediate memory, there exist regions of the maps that possess short memory and this short memory dominates.

For the polynomial map with $\alpha = 0.5$ the orbits possess 'weak' long memory, with

Estimated Standard Deviations of \hat{d} for $s_t^{(10)}$ series.							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.046*	0.070*	0.068*	0.013*	0.029*	0.012*	0.017*
Poly (0.45)	0.066*	0.088*	0.096*	0.034*	0.058*	0.026*	0.039*
Poly (0.5)	0.087*	0.112*	0.102*	0.045*	0.073*	0.030*	0.048*
Poly (0.65)	0.120*	0.156*	0.142*	0.054	0.139*	0.070*	0.068
Poly (0.8)	0.102*	0.136*	0.196*	0.098*	0.207*	0.117*	0.066*
Poly (0.9)	0.082*	0.111*	0.221*	0.133*	0.258*	0.147*	0.075*
Symmetric	0.101	0.123	0.160*	0.066*	0.117*	0.029*	0.059*
Asymmetric	0.101	0.125	0.178*	0.147*	0.133*	0.073*	0.078*
Log (0.05)	0.154	0.190	0.182*	0.112*	0.237	0.169	0.113
Log (0.15)	0.162	0.197	0.181*	0.119*	0.230	0.165	0.118
Log (0.25)	0.168	0.203	0.171*	0.113*	0.205	0.149	0.121
Log (0.3)	0.174	0.204	0.163*	0.119*	0.222	0.154	0.127
Estimated Standard Deviations of \hat{d} for $s_t^{(100)}$ series.							
Map	Local Whittle	GPH	FARIMA(AIC)	FARIMA(BIC)	FAR(AIC)	FAR(BIC)	FEXP
Poly (0.3)	0.045*	0.070*	0.055*	0.008*	0.020*	0.008*	0.016*
Poly (0.45)	0.052*	0.077*	0.054*	0.011*	0.030*	0.015*	0.025*
Poly (0.5)	0.052*	0.076*	0.063*	0.016*	0.029*	0.013*	0.025*
Poly (0.65)	0.110*	0.138*	0.106*	0.053*	0.096*	0.053*	0.070
Poly (0.8)	0.122*	0.155*	0.169*	0.097*	0.201*	0.119*	0.093*
Poly (0.9)	0.099*	0.129	0.190*	0.121*	0.239*	0.130*	0.093
Symmetric	0.084*	0.100*	0.089*	0.029*	0.063*	0.036*	0.052*
Asymmetric	0.080*	0.094*	0.097*	0.034*	0.070*	0.037*	0.051*
Log (0.05)	0.165	0.205	0.178*	0.139*	0.227	0.166	0.129
Log (0.15)	0.171	0.207	0.164*	0.132*	0.211	0.146	0.134
Log (0.25)	0.177	0.209	0.151*	0.118*	0.198	0.144	0.139
Log (0.3)	0.173	0.204	0.154*	0.105*	0.185	0.138	0.131

Tab. 4.6: The sample standard deviations for the estimates for the $s_t^{(10)}$ and $s_t^{(100)}$ series.

a singularity in the spectral density despite a d value of zero. It is suggested that the positive bias of the estimates of d for this map are caused by the inapplicability of the estimation techniques to these type of boundary long memory behaviour and estimates do seem to improve as the series length increases. This is discussed further in Chapter 5. For the polynomial map with $\alpha > 0.5$, the empirical evidence generally seems to agree with the asymptotic theory that $d = 1 - 1/2\alpha$.

For the asymmetric and symmetric cusp maps, the asymptotic theory again suggests 'weak' long memory of the same type as that for the polynomial map with $\alpha = 0.5$. Although this may explain some of the positive bias for these maps, this bias tends to be much larger than for the polynomial map with $\alpha = 0.5$, suggesting another source of bias is also present. The simulation studies of Bhansali, Holland and Kokoska [2006] and Bhansali and Holland [2008b] show for finite samples the periodogram near zero has steeper gradient than allowed for by the theory, the relative frequencies near the fixed point are often higher than the theory suggests and the ACF is larger at the lower lags than the other chaotic intermittency maps. This suggests a slow convergence rate of the maps to the asymptotic properties, possibly due to higher correlations at lower lags. The extended length simulation study, see section 4.2.1, shows the estimates appear to tend towards zero at a slow rate, whilst the systematic sampling study, see section 4.2.2, shows removal of the short range correlations can improve the estimates significantly.

The logarithmic map produces orbits at the upper boundary of long memory, being stationary despite a value of $d = 0.5$. As with the lower boundary case, the estimation techniques applied are not designed for this situation which could account for the negative bias. This is also discussed further in Chapter 5. However, since the magnitude of this bias seems similar to the cusp maps, it suggests another cause of bias may also be present. The simulation studies of Bhansali, Holland and Kokoska [2006] and Bhansali and Holland [2008b] suggest the finite sample behaviour only partly agrees with asymptotic theory, but since explicit forms of the ACF and spectral density are unknown this is difficult to quantify. The extended length simulation study does give some suggestion that the asymptotic results may hold, but that the rate of convergence is very slow.

The distributions of the estimates in all cases were not normal for $T < 10^6$ and the variances were generally much larger than the asymptotic theory would suggest. However, as seen in the extended length simulation study, the variances for the local Whittle and GPII estimates were seen to decrease as the series length increased, although at a rate generally slower than suggested by the asymptotic theory, and the distributions appear to become more regular. This suggests for the local Whittle and GPII methods some form of the consistency and possible asymptotic normality of the estimates may still hold for deterministic non-linear non-Gaussian chaotic intermittency maps, although at a slower rate. For the polynomial map with $\alpha = 0.3$, the asymptotic variances indeed appeared to hold. However, the persistence of the bias term when boundary long memory behaviour is present suggests these methods are not suitable in these situations, see Chapter 5 for further discussion.

The presence of the laminar regions caused difficulties in the FARIMA and FAR methods, generally giving poorer performance compared to the other estimation methods. Often, the fitted model explained the long memory by a near non-stationary short memory component and a negative value of d . This resulted in heavy tailed or even bimodal distributions of the estimates. The systematic sampling study showed the distributions could also be improved in terms of bias, standard deviation and \sqrt{MSE} , without changing the frequency of observations in the laminar region and that the difficulties arise due to long periods of consecutive values in the laminar region. The presence of such behaviour in actual time series should be obvious to find and this serves as a warning against use of these methods in such a situation.

The FExp method, although also producing estimates with non-Gaussian and large standard deviations and biases, generally outperformed the other methods in terms of bias, standard deviation and \sqrt{MSE} . theoretical justification as to why this method did not suffer as much from the same difficulties as the FAR method requires further work, although it is suggested that the effect of the laminar region is more apparent in the time domain and thus has more noticeable effect on the methods based on a time domain representation of the series.

5. DUAL PARAMETER LONG MEMORY MODEL

The standard methods of estimating the long memory parameter of a series, $\{x_t\}$, discussed in section 2.3, assume the spectral density of the series, f_x to be of the form

$$f_x(\lambda) \sim \lambda^{-2d} B(\lambda) \quad \text{as } \lambda \rightarrow 0, \quad (5.1)$$

where $B(\lambda)$ is a positive constant or bounded function as $\lambda \rightarrow 0$. The frequency analysis of the chaotic intermittency maps carried out by Bhansali and Holland [2008b], as discussed in section 3.2, shows the assumption that $B(\lambda)$ is bounded at $\lambda = 0$ does not hold for the symmetric and asymmetric cusp maps, the logarithmic maps or the polynomial map with $\alpha = 0.5$. The spectral density of an orbit, w_t , from any of these maps can be generalised as

$$f_w(\lambda) \sim \lambda^{-2d} \left(\log \left(\frac{1}{\lambda} \right) \right)^\gamma B(\lambda) \quad \text{as } \lambda \rightarrow 0, \quad (5.2)$$

where γ is an additional parameter and $B(\lambda)$ is a positive constant or bounded function as $\lambda \rightarrow 0$ as before, see section 3.2. The function $\log(1/\lambda)$ is slowly varying but unbounded at $\lambda = 0$. Thus, the extra parameter, γ , effects the rate at which the spectrum approaches infinity at zero and hence influences the long memory behaviour of the process. This parameter, γ , may thus be considered a second long memory parameter.

In particular, when $d = 0$ and $\gamma > 0$, the spectrum will still tend to infinity as λ tends to zero, though at a logarithmic rate rather than the standard polynomial rate when $d \in (0, 0.5)$. The autocorrelations of the series will thus not be absolute summable and hence the series will still possess long memory. This is the case for orbits from the cusp maps and the polynomial map with $\alpha = 0.5$, which have, asymptotically, $d = 0$ and $\gamma = 1$, as the orbit length $n \rightarrow \infty$. This boundary behaviour of long memory can be referred to as ‘weak’ long memory. Note, when $d = 0$, if $\gamma = 0$ the spectrum is bounded above and below at $\lambda = 0$ and the process has short memory, whilst if $\gamma < 0$, the spectrum will tend to zero as $\lambda \rightarrow 0$, which could be considered a form of ‘weak’ intermediate memory. The behaviour at this boundary condition is thus decided by the value of γ .

Alternatively, the boundary condition when $d = 0.5$ is normally non-stationary. However, as mentioned in Bhansali and Holland [2008b], when $d = 0.5$ and $\gamma < 0$,

the series will still be stationary, as is the case for orbits from the logarithmic map, with asymptotic values of $d = 0.5$ and $\gamma = -1 - \beta$ as $n \rightarrow \infty$. This upper boundary condition is referred to as 'strong' long memory. For $\gamma \geq 0$, this upper boundary will be non-stationary as in the standard case.

Note, also, equation 5.2 includes the standard form given in equation 5.1 as a special case where $\gamma = 0$. For $d \in (0, 0.5)$, a value of $\gamma < 0$ reduces the amount of long memory, whilst a value of $\gamma > 0$ increases it.

Previous works by Martin and Eccleston [1992], Martin and Walker [1997] and Palma [2007], page 61, introduced processes which produced 'weak' long memory. Other than the logarithmic map, very few processes currently considered in the literature appear to possess upper boundary long memory. The process studied by Martin and Eccleston [1992] and Martin and Walker [1997] is defined by its correlation structure, which is such that for some constant A , where $0 < A \leq (2 \log(2))^{-1}$,

$$r(u) = \frac{A}{|u|}, \quad u \neq 0. \quad (5.3)$$

From Theorem 4.1 of Bhansali and Holland [2008b], the spectral density of this process near zero is of the form

$$f(\lambda) \sim \frac{A}{\pi} \log \left(\frac{1}{2 \sin(\lambda/2)} \right), \quad \lambda \rightarrow 0.$$

Comparison of this with equation 5.2 shows this process has values of $d = 0$ and $\gamma = 1$ and thus admits 'weak' long memory behaviour.

The process given by Palma [2007], page 61 is defined by the Wold expansion

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} \frac{\varepsilon_{t-j}}{j} \quad (5.4)$$

where $\{\varepsilon_t\}$ is white noise. Palma [2007] gives its ACF is of the form

$$R(u) = \frac{1}{|u|} + \frac{1}{|u|} \sum_{j=1}^{|u|} \frac{1}{j}, \quad u \neq 0$$

$$R(0) = 1 + \frac{\pi^2}{6},$$

and

$$R(u) \sim \frac{\log(u)}{u}, \quad \text{as } u \rightarrow \infty.$$

Application of Theorem 2.15 on page 188 of Zygmund [1988] then gives the spectral density of this process as

$$f(\lambda) \sim B \log \left(\frac{1}{\lambda} \right)^2, \quad \lambda \rightarrow 0,$$

for some bounded B . Comparison of this with equation 5.2 shows this process has values of $d = 0$ and $\gamma = 2$, also admitting ‘weak’ long memory behaviour.

A similar process considered in section 5.4 has been developed by Prof. R. Bhansali and is included here with his permission. It is defined by its spectral density

$$f(\lambda) = \log \left(2 \sin \left(\frac{\lambda}{2} \right) \right)^2, \quad \lambda \in [-\pi, \pi]. \quad (5.5)$$

The following theorem, also by Prof. R. Bhansali, gives the ACF of this process.

Theorem 5.0.1. *For a process with spectral density defined by equation 5.5, the ACF is defined by*

$$R(0) = \frac{2\pi^3}{3},$$

$$R(1) = 4\pi,$$

$$R(u+1) = \frac{u}{u+1} R(u) + 4\pi \frac{2u+1}{u(u+1)^2}, \quad u \geq 2,$$

and, for some bounded B , as $u \rightarrow \infty$,

$$R(u) \sim B \frac{\log(u)}{u}.$$

Similar to the process of Palma [2007], this has values of $d = 0$ and $\gamma = 2$ and thus admits ‘weak’ long memory.

Not considered by these authors are methods of estimating d and γ , and the potential bias of estimating d using standard estimation techniques when $\gamma \neq 0$. The results of Chapter 4 showed this bias could be substantial. Further study of these biases could thus be an important issue for consideration. In section 5.1, the bias created in estimating d for the GPII method when falsely assuming $\gamma = 0$ is given and the method is extended to include this extra parameter. Similar treatment is then given to the Local Whittle method of estimating d in section 5.2.

Section 5.3 introduces a new extension to the FARIMA model which allows for a second long memory parameter. This model allows for both ‘weak’ and ‘strong’ long memory behaviour and includes the FARIMA model as a special case.

5.1 Dual Parameter GPH Method

This section introduces an extension of the GPH method of estimating d , described in section 2.3.2, to include the extra long memory parameter γ . To avoid confusion, the new extended method will be referred to as the *dual parameter GPH* (DGPH) method whilst the GPH method previously presented will be referred to as the *standard GPH* method. After defining the DGPH method, Theorem 5.1.1 will extend the results of the asymptotic properties of the periodogram given by Robinson [1995a] for the case when $\gamma \neq 0$.

These properties will then be used in Theorem 5.1.3 to show the bias that occurs for the standard GPH estimates when $\gamma \neq 0$. Theorem 5.1.4 will then show under certain conditions the new DGPH estimates of d and γ are asymptotically consistent and normally distributed.

The standard GPH method of estimating d , described in section 2.3.2, estimates d from the linear regression equation

$$\log(I(\lambda_j)) = \log(B(0)) - d \log(\lambda_j^2) + \log\left(\frac{B(\lambda_j)}{B(0)}\right) + \log\left(\frac{I(\lambda_j)}{f_x(\lambda_j)}\right),$$

where $I(\lambda_j)$ is the periodogram function, with $\lambda_j = 2\pi j/n$, $j \in (1, \dots, m)$, $\log(B(\lambda_j)/B(0))$ is considered negligible, n is the series length and $m/n \rightarrow 0$, $m \rightarrow \infty$ as $n \rightarrow \infty$. This regression equation comes from assuming the spectral density to be of the form given in equation 5.1, or equivalently by assuming $\gamma = 0$ in equation 5.2, then taking logs.

Similarly, taking logs of equation 5.2 without this assumption on γ and substituting in the periodogram leads to the equation

$$\log(I(\lambda_j)) = \log(B(0)) - d \log(\lambda_j^2) + \gamma \log(\log(\frac{1}{\lambda_j})) + \log\left(\frac{B(\lambda_j)}{B(0)}\right) + \log\left(\frac{I(\lambda_j)}{f_x(\lambda_j)}\right). \quad (5.6)$$

This extends the linear regression equation to include the γ parameter. The DGPH method is then to use this equation to find \hat{d} and $\hat{\gamma}$ using standard linear regression techniques.

Now, let $\{x_t\}$ be an observed time series of length n from a stationary process with spectral density $f(\lambda)$. Then,

$$w_j = w(\lambda_j) = \frac{\sum_{t=1}^n x_t e^{it\lambda_j}}{\sqrt{2\pi n}},$$

is the discrete Fourier transform of the observed series and

$$I_j = I(\lambda_j) = w_j \bar{w}_j,$$

where the bar above the w_j represents the complex conjugate. Define v_j as

$$v_j = v(\lambda_j) = \frac{w_j}{B\lambda_j^{-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma}, \quad (5.7)$$

where B is a positive bounded constant. The following two assumptions now stated are direct extensions of those given by Robinson [1995a] whilst studying log periodogram regression in the standard case, and Theorem 5.1.1 below is an extension of Theorem 2 of Robinson [1995a], noting that the v_j given in that paper are equivalent to the v_j given in equation 5.7 with $\gamma = 0$.

Assumption 5.1.1. There exists $B \in (0, \infty)$, $d \in (-0.5, 0.5)$, $\gamma \in (-\infty, \infty)$, or $d = 0.5$ with $\gamma \in (-\infty, 0)$, and $\alpha \in (0, 2]$ such that

$$f(\lambda) = B\lambda^{-2d} \log\left(\frac{1}{\lambda}\right)^\gamma + O\left(\lambda^{\alpha-2d} \log\left(\frac{1}{\lambda}\right)^\gamma\right) \quad \text{as } \lambda \rightarrow 0,$$

Assumption 5.1.2. In a neighbourhood $(0, \varepsilon)$ of the origin, $f(\lambda)$ is differentiable and

$$|f'(\lambda)| = \left| \frac{df(\lambda)}{d\lambda} \right| = O\left(\lambda^{-2d-1} \log\left(\frac{1}{\lambda}\right)^\gamma\right) \quad \text{as } \lambda \rightarrow 0,$$

The following Theorem holds with no further assumptions required on x_t or $f(\lambda)$.

Theorem 5.1.1. *Let assumptions 5.1.1 and 5.1.2 hold. Then for any sequences of positive integers $j = j(n)$ and $k = k(n)$ such that $j > k$ and $j/n \rightarrow 0$ as $n \rightarrow \infty$*

$$(a) \quad E(v_j \bar{v}_j) = 1 + O\left(\frac{\log(j)}{j} + \left(\frac{j}{n}\right)^\alpha\right)$$

$$(b) \quad E(v_j^2) = O\left(\frac{\log(j)}{j}\right)$$

$$(c) \quad E(v_j \bar{v}_k) = O\left(\frac{\log(j)}{k}\right)$$

$$(d) \quad E(v_j v_k) = O\left(\frac{\log(j)}{k}\right)$$

Proof. The proof is a direct extension of that given for Theorem 2 in Robinson [1995a], with the substitution of the assumptions 5.1.1 and 5.1.2 given here and v_j defined by equation 5.7. The following identities will be of use during the proof.

For any functions $g(x)$ and $h(x)$,

$$\left| \int_a^b g(x)h(x)dx \right| \leq \max_{a \leq x \leq b} |g(x)| \int_a^b |h(x)| dx, \quad (5.8)$$

and

$$\int_a^b |g(x) - h(x)| dx \leq \int_a^b |g(x)| dx + \int_a^b |h(x)| dx. \quad (5.9)$$

For a function, $g(x)$, such that $g(x) \geq 0$ for all $x \in [a, c]$ and any constant $b \in (a, c)$, then

$$\int_a^b g(x) dx \leq \int_a^c g(x) dx \quad (5.10)$$

and

$$\int_b^c g(x) dx \leq \int_a^c g(x) dx. \quad (5.11)$$

For an even function, such that $g(x) = g(-x)$, and some real constants a, b and c ,

$$\int_{-b}^{-a} g(x - c) dx = \int_a^b g(x + c) dx.$$

The mean value theorem states, for a function $g(x)$, there exists some constants a, b and c , such that $a < b < c$ and

$$g'(b) = \frac{g(c) - g(a)}{c - a}. \quad (5.12)$$

Now, take

$$\eta_{jk} = k^{-1} \log(j) \lambda_j^{-d} \lambda_k^{-d} \log(1/\lambda_j)^{\gamma/2} \log(1/\lambda_k)^{\gamma/2}, \quad (5.13)$$

and let

$$K(\lambda) = \frac{1}{2\pi n} \left| \sum_{s=1}^n \sum_{t=1}^n e^{i(t-s)\lambda} \right|^2,$$

be proportional to Fejer's kernel, such that the following properties hold for this kernel, see for example Anderson [1971] and Bruckner, Bruckner and Thomson [1997].

$$K(\lambda) = K(-\lambda),$$

$$K(\lambda) > 0,$$

$$K(\lambda) \leq \frac{1}{n\lambda^2}, \text{ for all } 0 < |\lambda| < \pi, \quad (5.14)$$

also

$$\int_{-\pi}^{\pi} K(\lambda - \lambda_j) d\lambda = 1, \quad \forall j \in (1, \dots, m) \quad (5.15)$$

and

$$E(I_j) = \int_{-\pi}^{\pi} f(\lambda) K(\lambda - \lambda_j) d\lambda. \quad (5.16)$$

The proof of part a) is to first show that

$$E(w_j \bar{w}_j) - f(\lambda_j) = O(\eta_{jj}) \quad (5.17)$$

and then to note that

$$f(\lambda_j) - B\lambda_j^{-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma = O\left(\left(\frac{j}{n}\right)^\alpha \lambda_j^{-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma\right),$$

from assumption 5.1.1.

Now, substitution of equation 5.16 into the left hand side of equation 5.17 and making use of the integral given in equation 5.15 being equal to 1 gives

$$\begin{aligned} & \left(\int_{-\pi}^{\pi} (f(\lambda)) K(\lambda - \lambda_j) d\lambda \right) - f(\lambda_j) = \\ & \int_{-\pi}^{\pi} (f(\lambda) - f(\lambda_j)) K(\lambda - \lambda_j) d\lambda. \end{aligned} \quad (5.18)$$

Choose a positive constant ε near zero, such that $2\lambda_j < \varepsilon$. The integral in equation 5.18 is separated into the following components

$$\int_{-\pi}^{-\varepsilon} + \int_{-\varepsilon}^{-\lambda_j/2} + \int_{-\lambda_j/2}^{\lambda_j/2} + \int_{\lambda_j/2}^{2\lambda_j} + \int_{2\lambda_j}^{\varepsilon} + \int_{\varepsilon}^{\pi}.$$

Each of these will be proved to be $O(\eta_{jj})$, proving equation 5.17 and thus completing the proof of part a).

The identity given in equation 5.8 shows the component of the integral given in equation 5.18 over $(-\pi, \varepsilon) \cup (\varepsilon, \pi)$ is bounded by

$$\left| \int_{-\pi}^{-\varepsilon} + \int_{\varepsilon}^{\pi} \right| \leq \max_{|\lambda| \geq \varepsilon} K(\lambda - \lambda_j) \int_{-\pi}^{\pi} |f(\lambda) - f(\lambda_j)| d\lambda$$

which is $o(\eta_{jj})$ by Lemma 5.1.1.

Next, using identity 5.9 the integral over $(-\varepsilon, -\lambda_j/2)$ can be split to give

$$\left| \int_{-\varepsilon}^{-\lambda_j/2} \right| \leq \int_{-\varepsilon}^{-\lambda_j/2} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{-\varepsilon}^{-\lambda_j/2} K(\lambda - \lambda_j) d\lambda.$$

and the integral over $(2\lambda_j, \varepsilon)$ can be split to give

$$\left| \int_{2\lambda_j}^{\varepsilon} \right| \leq \int_{2\lambda_j}^{\varepsilon} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{2\lambda_j}^{\varepsilon} K(\lambda - \lambda_j) d\lambda.$$

These are both $o(\eta_{jj})$ by Lemma 5.1.2

Making use of the mean value theorem, equation 5.12, implies

$$f(\lambda) - f(\lambda_j) = f'(b)(\lambda - \lambda_j),$$

for some $b \in (\lambda, \lambda_j)$ and substitution of this into the component of the integral given in equation 5.18 over $(\lambda_j/2, 2\lambda_j)$, along with use of identity 5.8 gives the bound

$$\left| \int_{\lambda_j/2}^{2\lambda_j} \right| \leq \left(\max_{\lambda_j/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \int_{\lambda_j/2}^{2\lambda_j} |\lambda - \lambda_j| K(\lambda - \lambda_j) d\lambda. \quad (5.19)$$

Now,

$$\left(\max_{\lambda_j/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) = O(\lambda_j^{-1-2d} \log(1/\lambda_j)^\gamma) \quad (5.20)$$

from the form of $f'(\lambda)$ given in assumption 5.1.2 and, using the upper limit of $K(\lambda - \lambda_j)$ from equation 5.14,

$$\int_{\lambda_j/2}^{2\lambda_j} |\lambda - \lambda_j| K(\lambda - \lambda_j) d\lambda = O\left(\frac{1}{n} \int_{\lambda_j/2}^{2\lambda_j} |\lambda - \lambda_j|^{-1} \right) = O\left(\frac{\log(j)}{n}\right) \quad (5.21)$$

see Robinson [1995a], with reference to Zygmund [1988] and Robinson [1994b]. Substitution of equations 5.20 and 5.21 into equation 5.19 gives

$$\left| \int_{\lambda_j/2}^{2\lambda_j} \right| = O\left(\lambda_j^{-1-2d} \log(1/\lambda_j)^\gamma \frac{\log(j)}{n}\right) = O(\eta_{jj}).$$

To complete the proof of equation 5.17, note, from the form of $f(\lambda)$ and $f(\lambda_j)$ given in assumption 5.1.1

$$\begin{aligned} \int_{-\lambda_j/2}^{\lambda_j/2} |f(\lambda)| d\lambda &= \left[\lambda^{1-2d} \log \left(\left| \frac{1}{\lambda} \right| \right)^\gamma + O \left(\lambda^{1-2d} \log \left(\left| \frac{1}{\lambda} \right| \right)^\gamma \right) \right]_{-\lambda_j/2}^{\lambda_j/2} \\ &= O \left(\lambda_j^{1-2d} \log \left(\frac{1}{\lambda_j} \right)^\gamma \right) \end{aligned} \quad (5.22)$$

and

$$\int_{-\lambda_j/2}^{\lambda_j/2} |f(\lambda_j)| d\lambda = |f(\lambda_j)| \int_{-\lambda_j/2}^{\lambda_j/2} 1 d\lambda = O \left(\lambda_j^{1-2d} \log(1/\lambda_j)^\gamma \right). \quad (5.23)$$

Now, the use of the identity given in equation 5.8 gives

$$\left| \int_{-\lambda_j/2}^{\lambda_j/2} \right| \leq \max_{|\lambda| \leq \lambda_j/2} K(\lambda - \lambda_j) \int_{-\lambda_j/2}^{\lambda_j/2} (|f(\lambda)| + |f(\lambda_j)|) d\lambda$$

and substitution of the bound of $K(\lambda - \lambda_j)$ given in equation 5.14 and the bounds of the integral given in equations 5.22 and 5.23 gives this is

$$= O \left(\frac{1}{n\lambda_j} \left(\lambda_j^{-2d} \log(1/\lambda_j)^\gamma \right) \right) = O(\eta_{jj}),$$

This completes the proof of part a).

Proving parts b), c) and d) is equivalent to proving

$$E(w_j^2) = O(\eta_{jj}),$$

$$E(w_j \bar{w}_k) = O(\eta_{jk}),$$

and

$$E(w_j w_k) = O(\eta_{jk}),$$

respectively.

Let $D(\lambda)$ be Dirichlet's kernel,

$$D(\lambda) = \sum_{t=1}^n e^{it\lambda},$$

and

$$E_{jk}(\lambda) = \frac{1}{2\pi n} D(\lambda_j - \lambda) D(\lambda - \lambda_k). \quad (5.24)$$

The following properties hold for these functions, see, for example, Zygmund [1988], Robinson [1994b] and Bruckner, Bruckner and Thomson [1997]

$$|D(\lambda)| \leq \frac{2}{|\lambda|}, \quad (5.25)$$

$$\int_{-C\lambda_j}^{C\lambda_j} |D(\lambda)| d\lambda = O(\log j), \quad (5.26)$$

for $0 < C < \infty$,

$$\int_{-\pi}^{\pi} E_{j,-k}(\lambda) d\lambda = 0 \quad \text{for } 1 \leq j+k < n. \quad (5.27)$$

while

$$E(w_j^2) = \int_{-\pi}^{\pi} (f(\lambda)) E_{j,-j}(\lambda) d\lambda, \quad (5.28)$$

$$E(w_j \bar{w}_k) = \int_{-\pi}^{\pi} f(\lambda) E_{jk}(\lambda) d\lambda, \quad \text{for } 0 < k < j < n, \quad (5.29)$$

and

$$E(w(\lambda_j)w(\lambda_k)) = \int_{-\pi}^{\pi} f(\lambda) E_{j,-k}(\lambda) d\lambda, \quad \text{for } 0 < k < j < n. \quad (5.30)$$

Note, substitution of the upper limit of $D(\lambda)$, given in equation 5.25, into the definition of $E_{jk}(\lambda)$, given in equation 5.24, gives

$$|E_{jk}(\lambda)| = \frac{1}{2\pi n} |D(\lambda_j - \lambda)| |D(\lambda - \lambda_k)| \leq \frac{2}{\pi n |\lambda - \lambda_j| |\lambda - \lambda_k|} \quad (5.31)$$

$$\leq \frac{2}{\pi n} \frac{1}{\min(|\lambda - \lambda_j|, |\lambda - \lambda_k|)^2}$$

From the integral in equation 5.27 being equal to zero, equation 5.28 is equivalent to

$$E(w_j^2) = \int_{-\pi}^{\pi} (f(\lambda) - f(\lambda_j)) E_{j,-j}(\lambda) d\lambda.$$

Similar to the proof of part a), this integral is decomposed into

$$\int_{-\pi}^{-\varepsilon} + \int_{-\varepsilon}^{-2\lambda_j} + \int_{-2\lambda_j}^{-\lambda_j/2} + \int_{-\lambda_j/2}^{\lambda_j/2} + \int_{\lambda_j/2}^{2\lambda_j} + \int_{2\lambda_j}^{\varepsilon} + \int_{\varepsilon}^{\pi}$$

and each of these will be proved to be $O(\eta_{jj})$. The calculations are similar to those in part a) and thus presented in an abbreviated form.

$$\left| \int_{-\pi}^{-\varepsilon} + \int_{\varepsilon}^{\pi} \right| \leq \max_{|\lambda| \geq \varepsilon} E_{j,-j}(\lambda) \int_{-\pi}^{\pi} |f(\lambda) - f(\lambda_j)| d\lambda = o(\eta_{jj}) \quad (5.32)$$

The upper limit of $E_{j,-j}(\lambda)$ from equation 5.31 gives

$$|E_{j,-j}(\lambda)| \leq \frac{2}{\pi n |\lambda + \lambda_j|^2},$$

implying Lemma 5.1.2 is applicable to give

$$\left| \int_{-\varepsilon}^{-2\lambda_j} + \int_{2\lambda_j}^{\varepsilon} \right| = o(\eta_{jj}).$$

Making use of the mean value theorem,

$$\left| \int_{-2\lambda_j}^{-\lambda_j/2} + \int_{\lambda_j/2}^{2\lambda_j} \right| = O \left(\left(\max_{\lambda_j/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \int_{\lambda_j/2}^{2\lambda_j} |\lambda - \lambda_j| |E_{j,-j}(\lambda)| d\lambda \right). \quad (5.33)$$

$$= O \left(\left(\max_{\lambda_j/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \frac{1}{n} \int_{-3\lambda_j}^{3\lambda_j} |D(\lambda)| d\lambda \right) = O(\eta_{jj})$$

Finally,

$$\left| \int_{-\lambda_j/2}^{\lambda_j/2} \right| \leq \max_{|\lambda| \leq \lambda_j/2} E_{j,-j}(\lambda) \int_{-\lambda_j/2}^{\lambda_j/2} (|f(\lambda)| + |f(\lambda_j)|) d\lambda$$

$$= O \left(\frac{1}{n\lambda_j} \left(\lambda_j^{-2d} \log(1/\lambda_j)^\gamma \right) \right) = O(\eta_{jj}),$$

This completes the proof of part b).

The proof of part c), it is equivalent to proving the integral given in equation 5.29 is $O(\eta_{jk})$. Making use of the integral from equation 5.27 being zero, the integral of equation 5.29 is expanded to

$$\int_{(\lambda_j+\lambda_k)/2}^{2\lambda_j} (f(\lambda) - f(\lambda_j)) E_{jk}(\lambda) d\lambda \quad (5.34)$$

$$+ \int_{\lambda_k/2}^{(\lambda_j+\lambda_k)/2} (f(\lambda) - f(\lambda_k)) E_{jk}(\lambda) d\lambda \quad (5.35)$$

$$- (f(\lambda_j) - f(\lambda_k)) \int_{\lambda_k/2}^{(\lambda_j+\lambda_k)/2} E_{jk}(\lambda) d\lambda \quad (5.36)$$

$$+ \left(\int_{-\lambda_k/2}^{\lambda_k/2} + \int_{-\lambda_j}^{-\lambda_k/2} \right) (f(\lambda) - f(\lambda_j)) E_{jk}(\lambda) d\lambda. \quad (5.37)$$

$$+ \left(\int_{-\pi}^{-\lambda_j} + \int_{2\lambda_j}^{\pi} \right) (f(\lambda) - f(\lambda_j)) E_{jk}(\lambda) d\lambda. \quad (5.38)$$

Using the same arguments as in parts a) and b), use of Lemmas 5.1.1 and 5.1.2 show equation 5.38 is $o(\eta_{jk})$.

For equation 5.34, making use of the mean value theorem gives

$$\begin{aligned} & \int_{(\lambda_j+\lambda_k)/2}^{2\lambda_j} (f(\lambda) - f(\lambda_j)) E_{jk}(\lambda) d\lambda \\ &= O \left(\left(\max_{(\lambda_j+\lambda_k)/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \int_{(\lambda_j+\lambda_k)/2}^{2\lambda_j} |\lambda - \lambda_j| |E_{j,-j}(\lambda)| d\lambda \right) \\ &= \left(\max_{(\lambda_j+\lambda_k)/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \frac{1}{n} \int_{(\lambda_j+\lambda_k)/2}^{2\lambda_j} |D(\lambda - \lambda_k)| d\lambda = O(\eta_{jj}) = O(\eta_{jk}). \end{aligned}$$

Consider equation 5.35. Now, following Robinson [1995a], the cases when $\lim_{n \rightarrow \infty} j/k < \infty$ and $\lim_{n \rightarrow \infty} j/k \rightarrow \infty$ are treated separately. Firstly, when $\lim_{n \rightarrow \infty} j/k < \infty$, substitution of $f(\lambda) - f(\lambda_k) = f'(b) |\lambda - \lambda_k|$, for $b \in (\lambda, \lambda_j)$ gives

$$\left| \int_{\lambda_k/2}^{(\lambda_j+\lambda_k)/2} (f(\lambda) - f(\lambda_k)) E_{jk}(\lambda) d\lambda \right|$$

$$\begin{aligned}
&\leq \left(\max_{\lambda_k/2 \leq \lambda \leq (\lambda_j + \lambda_k)/2} |f'(\lambda)| \right) \frac{1}{n} \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |\lambda - \lambda_k| |E_{jk}(\lambda)| d\lambda, \\
&= \left(\max_{\lambda_k/2 \leq \lambda \leq (\lambda_j + \lambda_k)/2} |f'(\lambda)| \right) \frac{1}{n} \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |D(\lambda_j - \lambda)| d\lambda \\
&= O \left(\frac{\log j}{k} \lambda_k^{-2d} \log(1/\lambda_k)^\gamma \right).
\end{aligned}$$

Now, for $\lim_{n \rightarrow \infty} j/k < \infty$,

$$O \left(\frac{\log j}{k} \lambda_k^{-2d} \log(1/\lambda_k)^\gamma \right) = O \left(\frac{\log j}{k} \lambda_j^{-2d} \log(1/\lambda_j)^\gamma \right) = O(\eta_{jk}). \quad (5.39)$$

For $\lim_{n \rightarrow \infty} j/k \rightarrow \infty$, first note

$$\begin{aligned}
\int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |E_{jk}(\lambda)| d\lambda &= \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} \left| \frac{1}{2\pi n} D(\lambda_j - \lambda) D(\lambda - \lambda_k) \right| d\lambda \quad (5.40) \\
&\leq \left(\max_{\lambda_k/2 \leq \lambda \leq (\lambda_j + \lambda_k)/2} \left| \frac{1}{2\pi n} D(\lambda_j - \lambda) \right| \right) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |D(\lambda - \lambda_k)| d\lambda \\
&= O \left(\left(\frac{1}{n\lambda_j - n\lambda_k} \right) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |D(\lambda - \lambda_k)| d\lambda \right) \\
&= O \left((j - k)^{-1} \int_{-\lambda_j}^{\lambda_j} |D(\lambda)| d\lambda \right),
\end{aligned}$$

since $\lambda_j = 2\pi j/n$. Thus, using identity 5.9 gives

$$\begin{aligned}
&\left| \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} (f(\lambda) - f(\lambda_k)) E_{jk}(\lambda) d\lambda \right| \leq \\
&\left(\max_{\lambda_k/2 \leq \lambda \leq \lambda_j} |f(\lambda)| + |f(\lambda_k)| \right) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |E_{jk}(\lambda)| d\lambda \quad (5.41)
\end{aligned}$$

and substitution of equation 5.40, the form of $f(\lambda)$ given in assumption 5.1.1 and the integral given in equation 5.26 shows this is

$$\begin{aligned}
&= O \left(\left(\lambda_j^{-2d} \log(1/\lambda_j)^\gamma + \lambda_k^{-2d} \log(1/\lambda_k)^\gamma \right) (j-k)^{-1} \int_{-\lambda_j}^{\lambda_j} |D(\lambda)| d\lambda \right) \\
&= O \left(\lambda_k^{-2d} \log(1/\lambda_k)^\gamma (j-k)^{-1} \log(j) \right).
\end{aligned}$$

Now, for $\lim_{n \rightarrow \infty} j/k \rightarrow \infty$

$$\lambda_k^{-2d} \log(1/\lambda_k)^\gamma (j-k)^{-1} \log(j) \rightarrow \lambda_k^{-2d} \log(1/\lambda_k)^\gamma (j)^{-1} \log(j) \quad (5.42)$$

$$= O \left(\left(\frac{k}{j} \right)^{(1-d)} \eta_{jk} \right) = o(\eta_{jk}),$$

as $n \rightarrow \infty$. Thus equation 5.35 is $O(\eta_{jk})$.

Next, consider equation 5.36. Again looking at the case when $\lim_{n \rightarrow \infty} j/k < \infty$, substitution of $f(\lambda_j) - f(\lambda_k) = f'(b)(\lambda_j - \lambda_k)$ for some $b \in (\lambda_k, \lambda_j)$ gives

$$\begin{aligned}
&\left| (f(\lambda_j) - f(\lambda_k)) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} E_{jk}(\lambda) d\lambda \right| \\
&\leq (\lambda_j - \lambda_k) \left(\max_{\lambda_k \leq \lambda \leq \lambda_j} |f'(\lambda)| \right) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} |E_{jk}(\lambda)| d\lambda \\
&\dots \\
&= O \left(\left(\lambda_k^{-1-2d} \log(1/\lambda_k)^\gamma \right) \frac{(\lambda_j - \lambda_k)}{(j-k)} \int_{-\lambda_j}^{\lambda_j} |D(\lambda)| d\lambda \right) = O \left(\left(\lambda_k^{-1-2d} \log(1/\lambda_k)^\gamma \right) \frac{\log(j)}{n} \right),
\end{aligned}$$

which, for $\lim_{n \rightarrow \infty} j/k < \infty$, is $O(\eta_{jk})$.

For $\lim_{n \rightarrow \infty} j/k \rightarrow \infty$,

$$\begin{aligned}
&\left| (f(\lambda_j) - f(\lambda_k)) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} E_{jk}(\lambda) d\lambda \right| \\
&\leq (|f(\lambda_j)| + |f(\lambda_k)|) \int_{\lambda_k/2}^{(\lambda_j + \lambda_k)/2} E_{jk}(\lambda) d\lambda \\
&= O \left((|f(\lambda_j)| + |f(\lambda_k)|) \frac{1}{(j-k)} \int_{-\lambda_j}^{\lambda_j} |D(\lambda)| d\lambda \right) = O \left(\lambda_k^{-2d} \log(1/\lambda_k)^\gamma (j-k)^{-1} \log(j) \right),
\end{aligned}$$

which, for $\lim_{n \rightarrow \infty} j/k \rightarrow \infty$, is $O(\eta_{jk})$.

Finally, consider equation 5.37. Now, note, from the upper limit of E_{jk} given in equation 5.31,

$$\max_{-\lambda_k/2 \leq \lambda \leq \lambda_k/2} |E_{jk}(\lambda)| = O\left(\frac{1}{n\lambda_j\lambda_k}\right) \quad (5.43)$$

Using identity 5.8 with $g(\lambda) = E_{jk}(\lambda)$ and $h(\lambda) = |f(\lambda)| + |f(\lambda_j)|$, then substituting in the result given in equation 5.43 and the form of $f(\lambda)$ given in assumption 5.1.1, the component of the integral given in equation 5.37 over $(-\lambda_k/2, \lambda_k/2)$ is bounded by

$$O\left(\frac{1}{n\lambda_j\lambda_k} \int_{-\lambda_k/2}^{\lambda_k/2} (|f(\lambda)| + |f(\lambda_j)|) d\lambda\right) = O(\eta_{jk})$$

and, using the same arguments as those given for equation 5.41, the component of the integral given in equation 5.37 over $(-\lambda_j, -\lambda_k/2)$ is bounded by

$$\begin{aligned} & O\left(\max_{\lambda_k/2 \leq \lambda \leq \lambda_j} |f(\lambda)| \frac{1}{j} \int_{-\lambda_j}^{\lambda_j} |D(\lambda - \lambda_k)| d\lambda\right) \\ & = O\left(\frac{\log j}{j} (\lambda_j^{-2d} \log(1/\lambda_j)^\gamma + \lambda_k^{-2d} \log(1/\lambda_k)^\gamma)\right) = O(\eta_{jk}). \end{aligned}$$

This completes the proof for part c).

The proof for part d) is similar to that of part c). However, there is no longer the need to distinguish between close and distant j, k .

Using the same arguments as in the previous parts,

$$\left| \int_{-\pi}^{-\varepsilon} + \int_{\varepsilon}^{\pi} \right| = o(\eta_{jk}),$$

$$\left| \int_{-\varepsilon}^{-2\lambda_j} + \int_{2\lambda_j}^{\varepsilon} \right| = O(\eta_{jk}).$$

$$\left| \int_{-2\lambda_j}^{-\lambda_k/2} \right| = O\left(\left(\max_{\lambda_k/2 \leq \lambda \leq 2\lambda_j} |f(\lambda)| + |f(\lambda_j)|\right) \int_{-2\lambda_j}^{-\lambda_k/2} |E_{j,-k}(\lambda)| d\lambda\right) = O(\eta_{jk}).$$

$$\left| \int_{\lambda_k/2}^{\lambda_j/2} \right| = O\left(\left(\max_{\lambda_k/2 \leq \lambda \leq \lambda_j} |f(\lambda)|\right) \frac{1}{j} \int_{-\lambda_j}^{\lambda_j} |D(\lambda)| d\lambda\right) = O(\eta_{jk}).$$

$$\left| \int_{-\lambda_k/2}^{\lambda_k/2} \right| = O \left(\max_{|\lambda| \leq \lambda_k/2} E_{j,-k}(\lambda) \int_{-\lambda_k/2}^{\lambda_k/2} (|f(\lambda)| + |f(\lambda_j)|) d\lambda \right) = O(\eta_{jk}).$$

$$\left| \int_{\lambda_j/2}^{2\lambda_j} \right| = O \left(\left(\max_{\lambda_j/2 \leq \lambda \leq 2\lambda_j} |f'(\lambda)| \right) \int_{\lambda_j/2}^{2\lambda_j} |\lambda - \lambda_j| |E_{j,-k}(\lambda)| d\lambda \right) = O(\eta_{jk}).$$

This completes the proof. □

Two further assumptions are now required to prove the asymptotic bias of the standard GPII estimates when $\gamma \neq 0$ and the asymptotic distribution of the new DGPII estimates. These assumptions are identical to assumptions 5 and 6 of Robinson [1995a].

Assumption 5.1.3. The series, x_t , $t \in (1, 2, \dots)$ is a Gaussian process.

Assumption 5.1.4. As $n \rightarrow \infty$

$$\frac{m^{1/2} \log(m)}{l} + \frac{l(\log(n))^2}{m} + \frac{m^{1+1/2\alpha}}{n} \rightarrow 0$$

Now, let

$$u_j = \log \left(\frac{I(\lambda_j)}{f(\lambda_j)} \right) - \psi(1), \quad j \in (l, \dots, m)$$

where ψ is the digamma function.

Theorem 5.1.2. Under assumptions 5.1.1-5.1.4, as $n \rightarrow \infty$, there exists independent zero mean random variables, ε_j , with variance $\sigma^2 = \pi^2/6$, such that

$$u_j \rightarrow_d \varepsilon_j$$

for $j \in (l, \dots, m)$.

The proof of Robinson [1995a] requires only assumptions 5.1.3 and 5.1.4 and the properties of v_j given in Theorem 5.1.1 and thus still holds.

The following theorem makes use of this to give the bias of the standard GPII method when the assumption that $\gamma = 0$ is false. The properties of the DGPII estimates which also makes use of this are discussed afterwards.

Theorem 5.1.3. *Let x_t be a series of length n such that assumptions 5.1.1-5.1.4 hold with $d = d_0$ and $\gamma = \gamma_0$. Then the expected value of the GPH estimate \hat{d} of d is given by*

$$E(\hat{d}) = d_0 + \frac{\gamma_0}{\log(\frac{n}{2\pi m})} + O\left(\frac{1}{\log(\frac{2\pi m}{n}) \log(\log(\frac{n}{2\pi m}))}\right),$$

as $n \rightarrow \infty$, $m \rightarrow \infty$, $m/n \rightarrow 0$.

Proof. Let Y be a vector $(y_1, \dots, y_m)'$ of observations, X be a matrix of regressors and Z be an additional, unaccounted for, matrix of regressors such that

$$\hat{B} = (X'X)^{-1}X'Y.$$

Standard linear regression techniques give

$$E(\hat{B}) = B + (X'X)^{-1}X'Z\gamma \quad (5.44)$$

Now, for the GPH method of estimating d , X is an $(m-l) \times 2$ matrix, with first column filled with 1's and second column given by $(\log((2\pi l/n)^2), \dots, \log((2\pi m/n)^2))'$, Z is an $(m-l) \times 1$ vector $(\log(\log(n/2\pi l)), \dots, \log(\log(2\pi m/n)))'$ and B is a 2×1 vector $(B, -d_0)'$.

The residual vector U is given by $(u_1, \dots, u_m)'$. Theorem 5.1.2 shows the effect of replacing these with independent zero mean variables is asymptotically negligible as $n \rightarrow \infty$ under assumptions 5.1.1- 5.1.4.

Substitution into equation 5.44 gives

$$E(\hat{d}) = d_0 - \gamma \frac{\sum_{j=l}^m (\log(2\pi j/n) - \bar{L}_1)(\log(\log(n/2\pi m)) - \bar{L}_2)}{2 \sum_{j=l}^m (\log(2\pi j/n) - \bar{L}_1)^2} \quad (5.45)$$

where

$$\bar{L}_1 = \frac{1}{m-l+1} \sum_{j=l}^m \log(2\pi j/n)$$

and

$$\bar{L}_2 = \frac{1}{m-l+1} \sum_{j=l}^m \log(\log(n/2\pi j)).$$

It has been shown, see Phillips [2001], that, given a slowly varying function $L(j)$,

$$\frac{1}{m} \sum_{j=1}^m L(j) = L(m) + 2 \frac{(mL'(m))^2}{L(m)} - 2mL'(m) - m^2L''(m) + o(mL'(m) + m^2L''(m)) \quad (5.46)$$

as $m \rightarrow \infty$. Use of equation 5.46 gives

$$\frac{\sum_{j=l}^m (\log(\frac{2\pi j}{n}) - \bar{L}_1)(\log(\log(\frac{n}{2\pi j})) - \bar{L}_2)}{2 \sum_{j=l}^m (\log(\frac{2\pi j}{n}) - \bar{L}_1)^2} = \frac{-1}{\log(\frac{n}{2\pi m})} + O\left(\frac{1}{\log(\frac{2\pi m}{n}) \log(\log(\frac{n}{2\pi m}))}\right)$$

as $m \rightarrow \infty$. This completes the proof. \square

Remark 5.1.1. The bias term added to the GPII method of the estimated d when γ is ignored tends to zero at a logarithmic rate. Hence, the GPII method still gives an asymptotically unbiased estimate for d .

Remark 5.1.2. Standard linear regression theory suggests the asymptotic variance of the GPII estimate remains unchanged. Hence, the variance will decrease at a faster rate than the bias, causing problems in hypothesis testing.

The following theorem now gives the limiting distribution of the DGPII estimates \hat{d} and $\hat{\gamma}$. The proof follows directly from Phillips [2001], Theorem 5.1, and is thus omitted.

Theorem 5.1.4. *Let x_t be a series of length n . Let assumptions 5.1.1-5.1.4 hold, with $d = d_0$, $\gamma = \gamma_0$. Let \hat{d} and $\hat{\gamma}$ be the least squares estimates of d_0 and γ_0 respectively, obtained from the regression equation 5.6, with $j \in (l, \dots, m)$ and $m/n \rightarrow 0$, $m \rightarrow \infty$ as $n \rightarrow \infty$. Then*

$$\frac{\sqrt{m}}{|\log(2\pi m/n)|} \begin{bmatrix} 2(\hat{d} - d_0) \\ \frac{1}{|\log(2\pi m/n)|}(\hat{\gamma} - \gamma_0) \end{bmatrix} \rightarrow_d N\left(0, \frac{\pi^2}{12} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right)$$

Remark 5.1.3. Note that the asymptotic results of Theorems 5.1.3 and 5.1.4 are obtained from taking the limits as n and m tend to ∞ of the various summations which appear in standard linear regression results. Hence, for a fixed n and m , these summations can be evaluated exactly, to give the *exact theoretical* results. Figure 5.1 shows the asymptotic results are only a good approximation for time series of length $n > 10^6$ and thus in practise the exact theoretical results should be used.

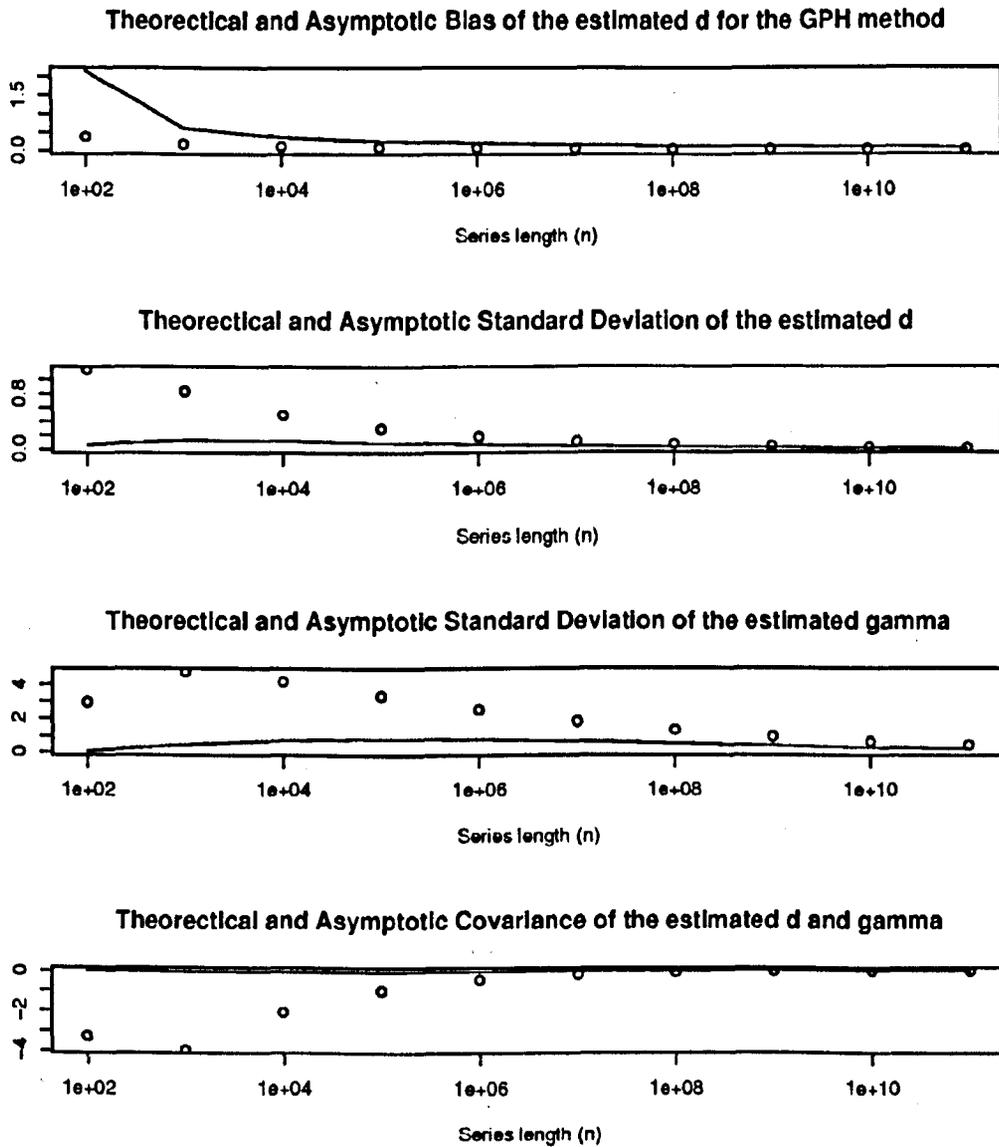


Fig. 5.1: The exact theoretical results, shown as dots, compared with the asymptotic results, shown as a line, for Theorems 5.1.3 and 5.1.4, with $m = \sqrt{n}$.

5.1.1 Technical Lemmas

This section contains technical lemmas required in the proof of Theorem 5.1.1.

Lemma 5.1.1. *Let $K(\lambda)$ be an a function such that $|K(\lambda)| \leq A/n\lambda^2$ for some constant A and $\lambda \neq 0$. Then, under assumptions 5.1.1 and 5.1.2, as $n \rightarrow \infty$,*

$$\max_{|\lambda| \geq \varepsilon} K(\lambda - \lambda_j) \int_{-\pi}^{\pi} |f(\lambda) - f(\lambda_j)| d\lambda = o(\eta_{jj})$$

Proof. Making use of identity 5.9 to separate this integral gives

$$\int_{-\pi}^{\pi} |f(\lambda) - f(\lambda_j)| d\lambda \leq \int_{-\pi}^{\pi} |f(\lambda)| d\lambda + \int_{-\pi}^{\pi} |f(\lambda_j)| d\lambda.$$

Now, since $\int_{-\pi}^{\pi} f(\lambda) d\lambda = R(0) < \infty$, substitution of the form of $f(\lambda_j)$ given in assumption 5.1.1, gives

$$\begin{aligned} \max_{|\lambda| \geq \varepsilon} K(\lambda - \lambda_j) \int_{-\pi}^{\pi} |f(\lambda) - f(\lambda_j)| d\lambda &\leq \left(\max_{|\lambda| \geq \varepsilon} K(\lambda - \lambda_j) \right) \times \\ &\left(R(0) + \left(B\lambda_j^{-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma + O\left(\lambda_j^{\alpha-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma\right) \right) \int_{-\pi}^{\pi} 1 d\lambda \right) \end{aligned}$$

then substituting the assumed upper limit of $K(\lambda - \lambda_j)$ gives that this is

$$\begin{aligned} &\leq \left(\frac{A}{n\varepsilon^2} \right) \left(R(0) + 2\pi B\lambda_j^{-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma + O\left(\lambda_j^{\alpha-2d} \log\left(\frac{1}{\lambda_j}\right)^\gamma\right) \right) \\ &= O\left(\frac{1 + \lambda_j^{-2d} \log(1/\lambda_j)^\gamma}{n} \right) = o(\eta_{jj}), \end{aligned}$$

where $\tilde{\varepsilon} = \varepsilon - \lambda_j > \lambda_j$ tends to a positive constant as $n \rightarrow \infty$, since

$$O\left(\frac{1}{n}\right) = O\left(\frac{(2\pi)^{2d} j^{1+2d}}{\lambda_j^{2d} j n^{1+2d}}\right) = O\left(\left(\frac{j}{n}\right)^{1+2d} \frac{\eta_{jj}}{\log(j) \log(1/\lambda_j)^\gamma n}\right) = o(\eta_{jj}),$$

for $d \in (-0.5, 0.5]$. This completes the proof. \square

Lemma 5.1.2. Let $K(\lambda)$ be an a function such that $|K(\lambda)| \leq A/n\lambda^2$ for some constant A and $\lambda \neq 0$. Then, under assumptions 5.1.1 and 5.1.2, as $n \rightarrow \infty$,

$$a) \int_{-\varepsilon}^{-\lambda_j/2} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{-\varepsilon}^{-\lambda_j/2} K(\lambda - \lambda_j) d\lambda = o(\eta_{jj})$$

and

$$b) \int_{2\lambda_j}^{\varepsilon} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{2\lambda_j}^{\varepsilon} K(\lambda - \lambda_j) d\lambda = o(\eta_{jj})$$

Proof. a) Making use of the identity given in equation 5.8 gives

$$\begin{aligned} & \left| \int_{-\varepsilon}^{-\lambda_j/2} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{-\varepsilon}^{-\lambda_j/2} K(\lambda - \lambda_j) d\lambda \right| \quad (5.47) \\ & \leq \left| \left(\max_{\lambda_j/2 \leq \lambda \leq \varepsilon} \frac{|f(\lambda)|}{\lambda^{(1-2d)/2}} \right) \int_{\lambda_j/2}^{\varepsilon} \lambda^{(1-2d)/2} K(-\lambda - \lambda_j) d\lambda \right. \\ & \quad \left. + |f(\lambda_j)| \int_{\lambda_j/2}^{\varepsilon} K(-\lambda - \lambda_j) d\lambda \right| \end{aligned}$$

Now, as $\lambda_j \rightarrow 0$,

$$\left(\max_{\lambda_j/2 \leq \lambda \leq \varepsilon} \frac{|f(\lambda)|}{\lambda^{(1-2d)/2}} \right) = O(\lambda_j^{-(1+2d)/2} \log(1/\lambda_j)^\gamma), \quad (5.48)$$

from the form of $f(\lambda)$ given in assumption 5.1.1. Also, substitution of the assumed upper limit of $K(\lambda)$ gives

$$\left| \int_{\lambda_j/2}^{\varepsilon} \lambda^{(1-2d)/2} K(-\lambda - \lambda_j) d\lambda \right| = O\left(\frac{1}{n} \int_{\lambda_j/2}^{\varepsilon} \lambda^{(-3-2d)/2} d\lambda\right)$$

and since $\lambda_j > 0$, identity 5.10 can be used on the upper limit of immigration to show this is

$$= O\left(\frac{1}{n} \int_{\lambda_j/2}^{\infty} \lambda^{(-3-2d)/2} d\lambda\right) = O\left(\frac{\lambda_j^{-(1+2d)/2}}{n}\right),$$

for $d \in (-0.5, 0.5]$, and similarly

$$\left| \int_{\lambda_j/2}^{\varepsilon} K(-\lambda - \lambda_j) d\lambda \right| = O\left(\frac{1}{n} \int_{\lambda_j/2}^{\infty} \lambda^{(-2)} d\lambda\right) = O\left(\frac{1}{j}\right), \quad (5.49)$$

since $\lambda_j = 2\pi j/n$. Substitution of equations 5.48 - 5.49 into equation 5.47 gives the required result.

b) Substituting $\bar{\lambda} = \lambda - \lambda_j$, an identical argument can be used to show

$$\begin{aligned} & \left| \int_{2\lambda_j}^{\varepsilon} |f(\lambda)| K(\lambda - \lambda_j) d\lambda + |f(\lambda_j)| \int_{2\lambda_j}^{\varepsilon} K(\lambda - \lambda_j) d\lambda \right| \\ &= O\left(\frac{1}{j} \lambda_j^{-2d} \log(1/\lambda_j)^\gamma\right) = o(\eta_{jj}). \end{aligned}$$

This completes the proof □

5.2 Dual Parameter Local Whittle Method

The Local Whittle method of estimating d , introduced and described in section 2.3.3, is extended in this section to include the extra long memory parameter γ . As with the extension of the GPII method, to avoid confusion the Local Whittle estimates will be referred to as the *standard Local Whittle* estimates, whilst the new extended method will be referred to as the *dual parameter Local Whittle* method. To motivate the need for this extension into dual parameters, Theorem 5.2.1 shows the bias of the standard Local Whittle estimate when $\gamma \neq 0$, although still $o(1)$, is of a higher order than the theoretical variance which from Robinson [1995b] is $O(1/m)$. The dual parameter Local Whittle estimates are then introduced and shown in Theorem 5.2.2 to be asymptotically consistent estimators of d_0 and γ_0 .

Now, let the standard Local Whittle estimate of d be given by

$$\hat{d} = \arg \min R(d)$$

with

$$R(d) = \log \hat{G}(d) - 2d \frac{1}{m} \sum_{j=1}^m \log(\lambda_j), \quad \hat{G}(d) = \frac{1}{m} \sum_{j=1}^m \lambda_j^{2d} I_j. \quad (5.50)$$

This is derived by substitution replacing $f(\lambda)$ by $G\lambda^{-2d}$ in the objective function

$$Q(G, d) = \frac{1}{m} \sum_{j=1}^m \left\{ \log f(\lambda_j) + \frac{I_j}{f(\lambda_j)} \right\} \quad (5.51)$$

where $I(\lambda)$ is the periodogram defined in equation 1.45 and the assumption is made that

$$f(\lambda) \sim G\lambda^{-2d} \quad \text{as } \lambda \rightarrow 0+, \quad (5.52)$$

that is, that $\gamma = 0$ in the dual parameter spectrum. The bias of these estimates when $\gamma \neq 0$ is found under the following assumptions.

Assumption 5.2.1. The observed series, $\{x_t\}$, is generated by a process with spectral density $f(\lambda)$, such that

$$f(\lambda) \sim G_0\lambda^{-2d_0} \log\left(\frac{1}{\lambda}\right)^{\gamma_0} \quad \text{as } \lambda \rightarrow 0+.$$

where $G_0 \in (0, \infty)$ and $(d_0, \gamma_0) \in \Theta = ((-0.5, 0.5), (-\infty, \infty)) \cup (0.5, (-\infty, 0))$.

Assumption 5.2.2. For $\lambda \in (0, \delta)$, $f(\lambda)$ is differentiable and

$$\frac{d}{d\lambda} \log(f(\lambda)) = O\left(\frac{1}{\lambda}\right),$$

as $\lambda \rightarrow 0+$.

Assumption 5.2.3. $\{x_t\}$ is a time series generated from a linear process, such that

$$x_t - E(x_0) = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} \alpha_j^2 < \infty,$$

where $E(\varepsilon_t | F_{t-1}) = 0$, $E(\varepsilon_t^2 | F_{t-1}) = 1$, a.s., $t = 0, \pm 1, \pm 2, \dots$, and F_t is the σ -field of events generated by ε_s , $s \leq t$. Also, there exists a random variable ε such that $E(\varepsilon^2) < \infty$ and for all $\nu > 0$ and some $K > 0$,

$$P(|\varepsilon_t| > \nu) \leq KP(|\varepsilon| > \nu)$$

Assumption 5.2.4. As $n \rightarrow \infty$, $m \rightarrow \infty$, but

$$\frac{\log(m)}{\log(n)} \rightarrow 0$$

Comparing these to the assumptions made in Robinson [1995b] for the standard Local Whittle estimate, it can be seen that Assumption 5.2.1 is a generalisation of assumption A1 in Robinson [1995b], allowing $\gamma \neq 0$, assumptions 5.2.2 and 5.2.3 are identical to assumptions A2 and A3 of Robinson [1995b] and assumption 5.2.4 is a stronger condition on the upper bound of m in relation to n . This may be stronger than required during the proof of the following theorem which shows the bias of the standard local Whittle estimate when the assumption that $\gamma_0 = 0$ does not hold,

but it is used later to prove the consistency of the new dual parameter local Whittle estimates.

Theorem 5.2.1. *Let $\{x_t\}$ be a series of length n generated by a linear process, such that assumptions 5.2.1- 5.2.4 hold.*

Let \hat{d} be the local Whittle estimate of d_0 found by minimising the function $R(d)$, given in equation 5.50, with respect to d . Then, for $\gamma \neq 0$, as $n \rightarrow \infty$,

$$0 <_p |\hat{d} - d_0| \leq_p \frac{|\gamma|}{2 \log(m)} \log \left(\frac{\log \left(\frac{n}{2\pi} \right)}{\log \left(\frac{n}{2\pi m} \right)} \right).$$

Proof. The derivative with respect to d of $R(d)$, defined in equation 5.50, is given by

$$R'(d) = \frac{2 \sum_{j=1}^m I_j \lambda_j^{2d} \log(\lambda_j)}{\sum_{j=1}^m \lambda_j^{2d} I_j} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j). \quad (5.53)$$

Now, Lemma 5.2.11 shows

$$R'(d) = \frac{-2 \sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma}}{\sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^\gamma} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1) \text{ as } n \rightarrow \infty, \quad (5.54)$$

and substitution of $d = d_0$ gives

$$R'(d_0) = \frac{-2 \sum_{j=1}^m \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma}}{\sum_{j=1}^m \log \left(\frac{1}{\lambda_j} \right)^\gamma} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1) \text{ as } n \rightarrow \infty.$$

By Lemma 5.2.12

$$R'(d_0) = -2\gamma + o_p(1) \text{ as } n \rightarrow \infty. \quad (5.55)$$

Thus, if $\gamma = 0$

$$R'(d_0) \rightarrow R'(\hat{d}) = 0, \text{ as } n \rightarrow \infty$$

implying

$$\hat{d} \rightarrow d_0, \text{ as } n \rightarrow \infty,$$

as proven by Robinson [1995b].

Lemma 5.2.1 shows the second derivative of $R(d)$ with respect to d , $R''(d)$, is positive for all $d \in \mathbb{R}$. This implies that the first derivative, $R'(d)$, increases with

d and thus if \hat{d} is defined such that $R'(\hat{d}) = 0$ then this point is an unique global minimum of $R(d)$.

Further, for $d > \hat{d}$

$$R'(d) > 0, \quad (5.56)$$

whilst for $d < \hat{d}$

$$R'(d) < 0. \quad (5.57)$$

Consider the case when $\gamma > 0$. The limit in probability of $R'(d_0)$ given in equation 5.55 is negative. From equation 5.57, this implies $d_0 < \hat{d}$ in probability.

For $\gamma > 0$ and $d > d_0$, as $n \rightarrow \infty$ equation 5.54 is bounded by

$$\begin{aligned} R'(d) &\geq \frac{\min \left(\lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^\gamma \right) 2 \sum_{j=1}^m \log(\lambda_j)}{\max \left(\lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^\gamma \right) \sum_{j=1}^m 1} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1) \quad (5.58) \\ &= m^{-2(d-d_0)} \left(\frac{\log \left(\frac{n}{2\pi} \right)}{\log \left(\frac{n}{2\pi m} \right)} \right)^\gamma \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1) \end{aligned}$$

Define d_1 such that

$$R'(d_1) \geq 0 + o_p(1)$$

as $n \rightarrow \infty$. Since $R'(d_1) \geq 0$ in probability, $d_1 \geq \hat{d}$ in probability. This gives, for $\gamma > 0$, as $n \rightarrow \infty$,

$$d_0 <_p \hat{d} \leq_p d_1$$

From equation 5.58 and the definition of d_1

$$m^{-2(d_1-d_0)} \left(\frac{\log \left(\frac{n}{2\pi} \right)}{\log \left(\frac{n}{2\pi m} \right)} \right)^\gamma = 1,$$

giving

$$d_1 = d_0 + \frac{|\gamma|}{2 \log(m)} \log \left(\frac{\log \left(\frac{n}{2\pi} \right)}{\log \left(\frac{n}{2\pi m} \right)} \right). \quad (5.59)$$

Now, define d_2 such that

$$R'(d_2) \leq 0 - o_p(1)$$

as $n \rightarrow \infty$. Since $R'(d_2) \leq 0$ in probability, $d_2 \leq \hat{d}$ in probability. This gives, for $\gamma < 0$, as $n \rightarrow \infty$,

$$d_2 \leq_p \hat{d} <_p d_0$$

Similar arguments as before gives

$$d_2 = d_0 - \frac{|\gamma|}{2 \log(m)} \log \left(\frac{\log \left(\frac{n}{2\pi} \right)}{\log \left(\frac{n}{2\pi m} \right)} \right). \quad (5.60)$$

Combining the cases of $\gamma > 0$ and $\gamma < 0$ gives

$$d_2 \leq_p \hat{d} \leq_p d_1$$

and substitution of d_1 and d_2 given in equations 5.59 and 5.60 gives the required result. This completes the proof. \square

The new dual parameter Local Whittle method is now introduced. The standard Local Whittle method is based on assumption 5.52 on the spectral density near zero. Here, this assumption is replaced with

$$f(\lambda) \sim G\lambda^{-2d} \log \left(\frac{1}{\lambda} \right)^\gamma \quad \text{as } \lambda \rightarrow 0+. \quad (5.61)$$

Substitution into equation 5.51 gives the dual parameter local Whittle estimates of d and γ as

$$(\hat{d}, \hat{\gamma}) = \arg \min R_2(d, \gamma)$$

with

$$R_2(d, \gamma) = \log \hat{G}(d, \gamma) - \frac{2d}{m} \sum_{j=1}^m \log(\lambda_j) + \frac{\gamma}{m} \sum_{j=1}^m \log(\log(1/\lambda_j)), \quad (5.62)$$

$$\hat{G}(d, \gamma) = \frac{1}{m} \sum_{j=1}^m \frac{\lambda_j^{2d} I_j}{\log(1/\lambda_j)^\gamma}. \quad (5.63)$$

The following Theorem shows the consistency of the dual parameter Local Whittle estimates of d_0 and γ_0 under the same assumptions given for Theorem 5.2.1 which showed the bias of the standard Local Whittle estimate of d_0 .

Theorem 5.2.2. *Let assumptions 5.2.1-5.2.4 hold.*

Let \hat{d} and $\hat{\gamma}$ be the dual parameter local Whittle estimates of d_0 and γ_0 respectively, found by minimising the function $R_2(d, \gamma)$ given in equation 5.62 with respect to d and γ . Then

$$\hat{d} \rightarrow_p d_0$$

and

$$\hat{\gamma} \rightarrow_p \gamma_0$$

as $n \rightarrow \infty$, $m \rightarrow \infty$, $m/n \rightarrow 0$.

Proof. The proof is an extension of that given by Robinson [1995b] for the standard Local Whittle estimate. For $0 < \delta < 0.5$, let $N_\delta = \{(d, \gamma) : |d - d_0| < \delta, |\gamma - \gamma_0| < \delta\} \cap \Theta$ and $\bar{N}_\delta = (\mathbb{R}^2 - N_\delta) \cap \Theta$. Define $S(d, \gamma) = R_2(d, \gamma) - R_2(d_0, \gamma_0)$.

Then

$$\begin{aligned} P\left(|\hat{d} - d_0| \geq \delta\right) &= P\left(\hat{d} \in \bar{N}_\delta\right) = P\left(\inf_{\bar{N}_\delta} R_2(d, \gamma) \leq \inf_{N_\delta} R_2(d, \gamma)\right) \\ &\leq P\left(\inf_{\bar{N}_\delta} S(d, \gamma) \leq 0\right) \end{aligned}$$

and

$$\begin{aligned} P(|\hat{\gamma} - \gamma_0| \geq \delta) &= P(\hat{\gamma} \in \bar{N}_\delta) = P\left(\inf_{\bar{N}_\delta} R_2(d, \gamma) \leq \inf_{N_\delta} R_2(d, \gamma)\right) \\ &\leq P\left(\inf_{\bar{N}_\delta} S(d, \gamma) \leq 0\right), \end{aligned}$$

because $(d_0, \gamma_0) \in N_\delta$.

Thus, if $P(\inf_{\bar{N}_\delta} S(d, \gamma) \leq 0) \rightarrow 0$ as $n \rightarrow \infty$ the proof will be complete. Let $\Theta_1 = ((d_0 - 0.5 + \eta, 0.5], (-\infty, \infty)) \cap \Theta$ and $\Theta_2 = \Theta - \Theta_1$, for some small positive constant η . Note, Θ_2 is empty if $d_0 \leq -\eta$. This gives

$$P\left(\inf_{\bar{N}_\delta} S(d, \gamma) \leq 0\right) = P\left(\inf_{\bar{N}_\delta \cap \Theta_1} S(d, \gamma) \leq 0\right) + P\left(\inf_{\bar{N}_\delta \cap \Theta_2} S(d, \gamma) \leq 0\right). \quad (5.64)$$

Consider the first probability on the right of equation 5.64. Separate $S(d, \gamma)$ into two parts such that

$$S(d, \gamma) = U(d) - T(d, \gamma),$$

where

$$\begin{aligned} T(d, \gamma) &= \log \left(\frac{\hat{G}(d_0, \gamma_0)}{G(d_0, \gamma_0)} \right) - \log \left(\frac{\hat{G}(d, \gamma)}{G(d, \gamma)} \right) - 2(d - d_0) \left(\frac{1}{m} \sum_{j=1}^m \log(j) - (\log(m) - 1) \right) \\ &\quad - \log \left(\frac{2(d - d_0) + 1}{m \log(1/\lambda_m)^{(\gamma_0 - \gamma)}} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d - d_0)} \log(1/\lambda_j)^{(\gamma_0 - \gamma)} \right) \\ &\quad + (\gamma - \gamma_0) \left(\frac{1}{m} \sum_{j=1}^m \log(\log(1/\lambda_j)) - \log(\log(1/\lambda_m)) \right), \\ U(d) &= 2(d - d_0) - \log(2(d - d_0) + 1), \end{aligned}$$

$$G(d, \gamma) = G_0 \frac{1}{m} \sum_{j=1}^m \lambda_j^{2(d - d_0)} \log(1/\lambda_j)^{\gamma_0 - \gamma}$$

and $\hat{G}(d, \gamma)$ is defined by equation 5.63.

The first probability on the right of equation 5.64 is bounded by

$$\therefore P \left(\sup_{\Theta_1} |T(d, \gamma)| \geq \inf_{N_\delta \cap \Theta_1} U(d) \right) \quad (5.65)$$

since $T(d, \gamma) \geq U(d)$ implies $S(d, \gamma) \leq 0$.

Note, Robinson [1995b] proves

$$\inf_{N_\delta \cap \Theta_1} U(d) > \frac{\delta^2}{2},$$

whilst Lemmas 5.2.13-5.2.16 show under assumptions 5.2.1-5.2.4 each of the terms of $T(d, \gamma)$ is $O_p(1/\log(\lambda_m))$ as $n \rightarrow \infty$. Thus the probability given in equation 5.65 tends to zero, implying the first probability on the right of equation 5.64 also tends to zero.

For the case $d_0 \leq -\eta$ the proof is complete. For $d_0 > -\eta$, the second probability in equation 5.64 must be considered. Take

$$p = p_m = \exp(m^{-1} \sum_{j=1}^m \log(j))$$

and $S(d, \gamma) = \log(\hat{D}(d, \gamma)/\hat{D}(d_0, \gamma_0))$, where

$$\hat{D}(d, \gamma) = \left(\frac{1}{m} \sum_{j=1}^m \log(1/\lambda_j)^{(\gamma-\gamma_0)} \right) \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p} \right)^{2(d-d_0)} j^{2d_0} \log(1/\lambda_j)^{-\gamma} I_j.$$

Lemma 5.2.17 shows

$$\begin{aligned} \hat{D}(d, \gamma) &= (1 + o(1)) \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p} \right)^{2(d-d_0)} j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j \\ &= (1 + o(1)) \tilde{D}(d), \end{aligned}$$

say. Using the identity $\log(1+x) \leq 2|x|$ for $|x| \leq 0.5$ gives

$$S(d, \gamma) \leq (1 + o(1)) \frac{\tilde{D}(d) - \tilde{D}(d_0)}{\tilde{D}(d_0)}.$$

The upper bound of this given in Lemma 5.2.18 shows

$$P \left(\inf_{\Theta_2} S(d, \gamma) \leq 0 \right) \leq P \left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j \leq 0 \right), \quad \text{as } n \rightarrow \infty,$$

where

$$a_j = \begin{cases} \left(\frac{j}{p} \right)^{(2\eta-1)}, & 1 \leq j \leq p \\ \left(\frac{j}{p} \right)^{2(-d_0-0.5)}, & p < j \leq m \end{cases}$$

and Lemma 5.2.19 shows this probability is in turn bounded by

$$P \left(\left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1 \right) \right| \geq 1 \right), \quad \text{as } n \rightarrow \infty. \quad (5.66)$$

Finally, Lemma 5.2.9 shows under assumptions 5.2.1-5.2.4, equation 5.66 is $o(1)$ as $n \rightarrow \infty$. Thus equation 5.64 is $o(1)$ as $n \rightarrow \infty$, as required. This completes the proof. \square

Robinson [1995b] strengthens his assumptions A1-A4 to a new set of assumptions A1'-A4', under which his Theorem 2 shows the asymptotic distribution of the standard Local Whittle estimate is normal. Theorem 5.2.3 below provides an extension of Theorem 2 given in Robinson [1995b] and is presented here under similar strengthening of assumptions 5.2.1-5.2.4. It is seen that, as with the dual parameter GPH estimates, the dual parameter local Whittle estimates are asymptotically collinear with a normal distribution.

Introduce the following assumptions,

Assumption 5.2.5. For some $\beta \in (0, 2]$,

$$f(\lambda) = G_0 \lambda^{-2d_0} \log\left(\frac{1}{\lambda}\right)^{\gamma_0} \left(1 + O(\lambda^\beta)\right) \quad \text{as } \lambda \rightarrow 0+.$$

where $G_0 \in (0, \infty)$ and $(d_0, \gamma_0) \in \Theta = ((-0.5, 0.5), (-\infty, \infty)) \cup (0.5, (-\infty, 0))$.

Assumption 5.2.6. Define $\alpha(\lambda)$ such that $f(\lambda) = |\alpha(\lambda)|^2 / 2\pi$. For $\lambda \in (0, \delta)$, $\alpha(\lambda)$ is differentiable and

$$\frac{d}{d\lambda} \alpha(\lambda) = O\left(\frac{\alpha(\lambda)}{\lambda}\right),$$

as $\lambda \rightarrow 0+$.

Assumption 5.2.7. Assumption 5.2.3 holds and also

$$E(\varepsilon_t^3 | I_{t-1}^1) = \mu_3, E(\varepsilon_t^4 | I_{t-1}^1) = \mu_4, \text{ a.s., } t = 0, \pm 1, \pm 2, \dots,$$

for finite constants μ_3 and μ_4 .

Assumption 5.2.8. There exists some $\nu > 0$ such that as $n \rightarrow \infty$,

$$\frac{\log(m)^{6+2\nu}}{\log(n)} + \frac{\log(\log(n))^2}{\log(m)^\nu} \rightarrow 0$$

Note, under assumptions 5.2.5-5.2.8, assumptions 5.2.1-5.2.4 still hold. Assumption 5.2.5 strengthens assumption 5.2.1 by giving a rate of convergence and is a generalisation of assumption A1' of Robinson [1995b], allowing $\gamma \neq 0$. Assumptions 5.2.6 and 5.2.7 are identical to assumptions A2' and A3' of Robinson [1995b]. Assumption 5.2.8 is a stronger upper bound on the rate at which $m \rightarrow \infty$ as $n \rightarrow \infty$ than given in assumption A4' of Robinson [1995b], which is of use during the proof of Theorem 5.2.3.

Theorem 5.2.3. Let \hat{d} and $\hat{\gamma}$ be the dual parameter Local Whittle estimates of d_0 and γ_0 found by minimising the objective function given in equation 5.50. Let assumptions 5.2.5-5.2.8 hold. Then, as $n \rightarrow \infty$,

$$\frac{2m^{1/2}}{\log(n)} \begin{pmatrix} 2(\hat{d} - d_0) \\ \frac{1}{\log(n)}(\hat{\gamma} - \gamma_0) \end{pmatrix} \rightarrow_d N \left(0, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right).$$

Proof. Let $R_d(d, \gamma)$ be the partial differential of $R_2(d, \gamma)$ with respect to d , $R_\gamma(d, \gamma)$ be the partial differential of $R_2(d, \gamma)$ with respect to γ , and $R_{dd}(d, \gamma), R_{d\gamma}(d, \gamma), R_{\gamma\gamma}(d, \gamma)$ be the partial differentials of these with respect to d and γ respectively. Then, Theorem 5.2.2 implies the Taylor expansions

$$0 = R_d(\hat{d}, \hat{\gamma}) = R_d(d_0, \gamma_0) + R_{dd}(\bar{d}, \bar{\gamma})(\hat{d} - d_0) + R_{d\gamma}(\bar{d}, \bar{\gamma})(\hat{\gamma} - \gamma_0) \quad (5.67)$$

and

$$0 = R_\gamma(\hat{d}, \hat{\gamma}) = R_\gamma(d_0, \gamma_0) + R_{\gamma d}(\bar{d}, \bar{\gamma})(\hat{d} - d_0) + R_{\gamma\gamma}(\bar{d}, \bar{\gamma})(\hat{\gamma} - \gamma_0) \quad (5.68)$$

both hold, where $|\bar{d} - d_0| < |\hat{d} - d_0|$ and $|\bar{\gamma} - \gamma_0| < |\hat{\gamma} - \gamma_0|$.

Note, under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$, Lemma 5.2.22 shows

$$R_{dd}(\bar{d}, \bar{\gamma}) = 4 + o_p \left(\frac{1}{\log(m)} \right),$$

Lemma 5.2.23 shows

$$R_{d\gamma}(\bar{d}, \bar{\gamma}) = \frac{-2}{\log(2\pi m/n)} - \frac{4}{\log(2\pi m/n)^2} + o_p \left(\frac{1}{\log(2\pi m/n)^2 \log(\log(n/(2\pi m)))} \right),$$

Lemma 5.2.24 shows

$$R_{\gamma\gamma}(\bar{d}, \bar{\gamma}) = \frac{1}{\log(2\pi m/n)^2} + \frac{4}{\log(2\pi m/n)^3} + o_p \left(\frac{1}{\log(2\pi m/n)^3} \right),$$

and Lemmas 5.2.25, 5.2.27 and 5.2.29 show

$$\frac{m^{1/2}}{2} R_d(d_0, \gamma_0) \rightarrow_d X$$

and

$$\log(2\pi m/n) m^{1/2} R_\gamma(d_0, \gamma_0) \rightarrow_d Y$$

where

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(0, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right),$$

Substitution of these into equations 5.67 and 5.68 gives

$$\begin{pmatrix} 1 & -\left(1 + \frac{2}{\log(\lambda_m)}\right) \\ -\left(1 + \frac{2}{\log(\lambda_m)}\right) & \left(1 + \frac{4}{\log(\lambda_m)}\right) \end{pmatrix} \begin{pmatrix} 2m^{1/2}(\hat{d} - d_0) \\ \frac{(m)^{1/2}}{\log(\lambda_m)}(\hat{\gamma} - \gamma_0) \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix} + o_p(1),$$

which can be written as

$$\frac{2m^{1/2}}{\log(\lambda_m)} \begin{pmatrix} \frac{\log(\lambda_m)}{2} & -\left(\frac{\log(\lambda_m)}{2} + 1\right) \\ -\left(\frac{\log(\lambda_m)}{2} + 1\right) & \left(\frac{\log(\lambda_m)}{2} + 2\right) \end{pmatrix} \begin{pmatrix} 2(\hat{d} - d_0) \\ \frac{1}{\log(\lambda_m)}(\hat{\gamma} - \gamma_0) \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix} + o_p(1). \quad (5.69)$$

Let A be the 2×2 matrix in equation 5.69. Standard matrix inversion techniques give

$$\det A = \frac{\log(\lambda_m)^2}{4} + \log(\lambda_m) - \frac{\log(\lambda_m)^2}{4} - \log(\lambda_m) + 1 = 1$$

and

$$A^{-1} = \begin{pmatrix} \left(\frac{\log(\lambda_m)}{2} + 2\right) & \left(\frac{\log(\lambda_m)}{2} + 1\right) \\ \left(\frac{\log(\lambda_m)}{2} + 1\right) & \left(\frac{\log(\lambda_m)}{2}\right) \end{pmatrix}.$$

Hence,

$$\frac{2m^{1/2}}{\log(\lambda_m)} \begin{pmatrix} 2(\hat{d} - d_0) \\ \frac{1}{\log(\lambda_m)}(\hat{\gamma} - \gamma_0) \end{pmatrix} = \begin{pmatrix} \left(\frac{\log(\lambda_m)}{2} + 2\right) & \left(\frac{\log(\lambda_m)}{2} + 1\right) \\ \left(\frac{\log(\lambda_m)}{2} + 1\right) & \left(\frac{\log(\lambda_m)}{2}\right) \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + o_p(1).$$

$$\rightarrow_d N(0, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} \left(\frac{\log(\lambda_m)}{2} + 2\right) & \left(\frac{\log(\lambda_m)}{2} + 1\right) \\ \left(\frac{\log(\lambda_m)}{2} + 1\right) & \left(\frac{\log(\lambda_m)}{2}\right) \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \left(\frac{\log(\lambda_m)}{2} + 2\right) & \left(\frac{\log(\lambda_m)}{2} + 1\right) \\ \left(\frac{\log(\lambda_m)}{2} + 1\right) & \left(\frac{\log(\lambda_m)}{2}\right) \end{pmatrix}$$

$$\begin{aligned}
&= \left(\begin{array}{cc} \left(\frac{\log(\lambda_m)}{2} + 2 \right) & \left(\frac{\log(\lambda_m)}{2} + 1 \right) \\ \left(\frac{\log(\lambda_m)}{2} + 1 \right) & \left(\frac{\log(\lambda_m)}{2} \right) \end{array} \right) \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},
\end{aligned}$$

as required. This completes the proof. \square

5.2.1 Technical Lemmas

The Lemmas required for the proofs of Theorems 5.2.1, 5.2.2 and 5.2.3 are listed in this section.

Lemma 5.2.1. *The second derivative with respect to d of $R(d)$ defined in equation 5.50 is positive for all $d \in \mathbb{R}$.*

Proof. From the first derivative of $R(d)$, given in equation 5.53, the second derivative of $R(d)$ with respect to d is

$$R''(d) = \frac{4 \sum_{j=1}^m \sum_{k=1}^m I_j I_k \lambda_j^{2d} \lambda_k^{2d} (\log(\lambda_j)^2 - \log(\lambda_j) \log(\lambda_k))}{\left(\sum_{j=1}^m \lambda_j^{2d} I_j \right)^2}$$

Due to the symmetry in j and k , this can be written as

$$\begin{aligned}
&= \frac{2 \sum_{j=1}^m \sum_{k=1}^m I_j I_k \lambda_j^{2d} \lambda_k^{2d} (\log(\lambda_j)^2 - 2 \log(\lambda_j) \log(\lambda_k) + \log(\lambda_k)^2)}{\left(\sum_{j=1}^m \lambda_j^{2d} I_j \right)^2} \\
&= \frac{2 \sum_{j=1}^m \sum_{k=1}^m I_j I_k \lambda_j^{2d} \lambda_k^{2d} (\log(\lambda_j) - \log(\lambda_k))^2}{\left(\sum_{j=1}^m \lambda_j^{2d} I_j \right)^2}. \tag{5.70}
\end{aligned}$$

Since $I_j, \lambda_j > 0$ for $j \in (1, \dots, m)$ and the bracketed terms in both numerator and denominator are squared, equation 5.70 implies $R''(d) > 0$ for any real value of d . This completes the proof. \square

Define g_j such that

$$g_j = G_0 \lambda_j^{-2d_0} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0} \tag{5.71}$$

and write $I_j/g_j - 1$ as

$$\frac{I_j}{g_j} - 1 = \left(1 - \frac{g_j}{f_j}\right) \frac{I_j}{g_j} + \frac{1}{f_j} \left(I_j - |\alpha_j|^2 I_{\varepsilon j}\right) + (2\pi I_{\varepsilon j} - 1),$$

where $I_{\varepsilon j} = I_{\varepsilon} = |w_{\varepsilon}(\lambda_j)|^2$, $w_{\varepsilon}(\lambda) = (2\pi n)^{-0.5} \sum_{t=1}^n \varepsilon_t e^{it\lambda}$, $f_j = f(\lambda_j)$ and $\alpha_j = \alpha(\lambda_j) = \sum_{l=0}^{\infty} \alpha_l e^{il\lambda_j}$. The properties given in Lemmas 5.2.2-5.2.4 are generalisations of the properties proven by Robinson [1995b] when γ_0 can be taken as nonzero.

Lemma 5.2.2. *Under assumptions 5.2.1- 5.2.4, for any $\nu > 0$,*

$$\left|1 - \frac{g_j}{f_j}\right| \leq \nu, \quad \text{as } n \rightarrow \infty, \quad \forall j \in (1, \dots, m).$$

Proof. From assumption 5.2.1

$$f_j = g_j + o(g_j),$$

which gives

$$\left|1 - \frac{g_j}{f_j}\right| = |1 - 1 + o(1)| = o(1),$$

as required. □

Lemma 5.2.3. *Under assumptions 5.2.1- 5.2.4, for any $\nu > 0$,*

$$E \left|1 - \frac{g_j}{f_j}\right| \left|\frac{I_j}{g_j}\right| \leq \nu, \quad \text{as } n \rightarrow \infty, \quad \forall j \in (1, \dots, m).$$

Proof. Under assumptions 5.2.1- 5.2.4, Lemma 5.2.2 and Theorem 5.1.1 shows that

$$E \left|1 - \frac{g_j}{f_j}\right| \left|\frac{I_j}{g_j}\right| < \tilde{\nu} \left(1 + A_j \frac{\log(j+1)}{j}\right) < \nu, \quad \text{as } n \rightarrow \infty, \quad \forall j \in (1, \dots, m),$$

for some bounded $|A_j| < \infty$, $\forall j \in (1, \dots, m)$, as required. □

Lemma 5.2.4. *Under assumptions 5.2.1- 5.2.4,*

$$\frac{1}{f_j} E \left|I_j - |\alpha_j|^2 I_{\varepsilon j}\right| = O \left(\left(\frac{\log(j+1)}{j} \right)^{1/2} \right), \quad \text{as } n \rightarrow \infty,$$

Proof. Under assumptions 5.2.1- 5.2.4, the proof of Theorem 5.1.1 shows that

$$E(I_j) = g_j \left(1 + A_j \frac{\log(j+1)}{j} \right), \quad (5.72)$$

$$\bar{\alpha}_j E(w_j \bar{w}_{\varepsilon_j}) = g_j \left(1 + B_j \frac{\log(j+1)}{j} \right),$$

$$\alpha_j E(\bar{w}_j w_{\varepsilon_j}) = g_j \left(1 + C_j \frac{\log(j+1)}{j} \right),$$

and

$$|\alpha_j|^2 E(I_{\varepsilon_j}) = g_j \left(1 + D_j \frac{\log(j+1)}{j} \right), \quad (5.73)$$

for bounded $|A_j|, |B_j|, |C_j|, |D_j| < \infty, \forall j \in (1, \dots, m)$ as $n \rightarrow \infty$. The remainder of the proof then follows directly from Robinson [1995b] and is thus omitted. \square

Use of Lemmas 5.2.2-5.2.4 shows the proofs of Robinson [1995b] still hold to give the following results of Lemmas 5.2.5 - 5.2.9

Lemma 5.2.5. *Under assumptions 5.2.1-5.2.4, as $n \rightarrow \infty$,*

$$\frac{1}{m} \sum_{j=1}^m \left| \left(\frac{I_j}{g_j} - 1 \right) \left(\frac{j}{m} \right)^{2(d-d_0)} \right| = O_p \left(\frac{1}{m^{1/2}} \right).$$

Lemma 5.2.6. *Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$E \left| \sum_{j=1}^m \left(\frac{I_j}{g_j} - 1 \right) \right| = O_p \left(m^{1/2} + (m \log(m)^2)^{1/3} + \frac{m^{\beta+1}}{n^\beta} + \frac{m^{1/2}}{n^{1/4}} \right).$$

Lemma 5.2.7. *Under assumptions 5.2.5-5.2.8, for v_j defined such that, as $n \rightarrow \infty$,*

$$\sum_{j=1}^m v_j = 0,$$

$$\frac{1}{m^{1/2}n} \sum_{j=1}^m |v_j| = o(1)$$

and

$$\frac{1}{m} \sum_{j=1}^m v_j^2 = 1 + o(1),$$

then, as $n \rightarrow \infty$,

$$\frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \left(\frac{I_j}{g_j} - 1 \right) \right) \rightarrow_d N(0, 1)$$

Lemma 5.2.8. Under assumptions 5.2.5-5.2.8, for v_j and η_j defined such that, as $n \rightarrow \infty$,

$$\frac{1}{m} \sum_{j=1}^m v_j \eta_j = 1 + o(1),$$

then, as $n \rightarrow \infty$,

$$E \left\{ \left(\frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \left(\frac{I_j}{g_j} - 1 \right) \right) \right) \left(\frac{1}{m^{1/2}} \left(\sum_{j=1}^m \eta_j \left(\frac{I_j}{g_j} - 1 \right) \right) \right) \right\} \rightarrow 1$$

Lemma 5.2.9. Under assumptions 5.2.1-5.2.4, for a small positive constant η , with $d_0 > -\eta$,

$$p = p_m = \exp(m^{-1} \sum_{j=1}^m \log(j))$$

and

$$a_j = \begin{cases} \left(\frac{j}{p} \right)^{(2\eta-1)}, & 1 \leq j \leq p \\ \left(\frac{j}{p} \right)^{2(-d_0-0.5)}, & p < j \leq m \end{cases}$$

then as $n \rightarrow \infty$

$$P \left(\left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1 \right) \right| \geq 1 \right) = o(1).$$

Lemma 5.2.10. Under assumptions 5.2.1- 5.2.4, as $n \rightarrow \infty$,

$$\frac{1}{m} \sum_{j=1}^m I_j \lambda_j^{2d} \log(\lambda_j)^b = (-1)^b \left(G_0 \left(\frac{2\pi m}{n} \right)^{2(d-d_0)} \right) \left(\frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{b+\gamma_0} \right)$$

$$+O_p\left(\frac{1}{m^{1/2}} \max_{j \in (1, \dots, m)} \left| \log\left(\frac{1}{\lambda_j}\right)^{b+\gamma_0} \right| \right),$$

for some integer b .

Proof. Take g_j as defined in equation 5.71. Then

$$\frac{1}{m} \sum_{j=1}^m I_j \lambda_j^{2d} \log(\lambda_j)^b = \frac{1}{m} \sum_{j=1}^m \frac{g_j}{g_j} I_j \lambda_j^{2d} \log(\lambda_j)^b, \quad (5.74)$$

$$= (-1)^b \frac{1}{m} \sum_{j=1}^m \frac{I_j}{g_j} G_0 \lambda_j^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0}$$

$$= (-1)^b \frac{1}{m} \sum_{j=1}^m \left(\frac{I_j}{g_j} - 1 \right) G_0 \lambda_j^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0} + (-1)^b \frac{1}{m} \sum_{j=1}^m G_0 \lambda_j^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0}$$

and substitution of $\lambda_j = 2\pi j/m$ gives this is

$$\begin{aligned} &= (-1)^b \left(G_0 \left(\frac{2\pi m}{n} \right)^{2(d-d_0)} \right) \left(\frac{1}{m} \sum_{j=1}^m \left(\frac{I_j}{g_j} - 1 \right) \left(\frac{j}{m} \right)^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0} \right. \\ &\quad \left. + \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0} \right) \\ &= (-1)^b \left(G_0 \left(\frac{2\pi m}{n} \right)^{2(d-d_0)} \right) (A_1(d) + A_2(d)), \end{aligned}$$

say.

Consider $A_1(d)$.

$$\begin{aligned} |A_1(d)| &= \frac{1}{m} \sum_{j=1}^m \left| \left(\frac{I_j}{g_j} - 1 \right) \left(\frac{j}{m} \right)^{2(d-d_0)} \log(1/\lambda_j)^{b+\gamma_0} \right| \\ &\leq \max_{j \in (1, \dots, m)} \left| \log(1/\lambda_j)^{b+\gamma_0} \right| \frac{1}{m} \sum_{j=1}^m \left| \left(\frac{I_j}{g_j} - 1 \right) \left(\frac{j}{m} \right)^{2(d-d_0)} \right| \end{aligned}$$

Under assumptions 5.2.1- 5.2.4, Lemma 5.2.5 can be used to give

$$|A_1(d)| = O_p \left(\frac{1}{m^{1/2}} \max_{j \in (1, \dots, m)} \left| \log(1/\lambda_j)^{b+\gamma_0} \right| \right), \text{ as } n \rightarrow \infty. \quad (5.75)$$

Substitution of equation 5.75 into equation 5.74 gives the required result. \square

Lemma 5.2.11. *Under assumptions 5.2.1- 5.2.4, as $n \rightarrow \infty$, the first derivative with respect to d of $R(d)$ defined in equation 5.50 is equal to*

$$R'(d) = \frac{-2 \sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma}}{\sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^\gamma} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1).$$

Proof. Write $R'(d)$, given in equation 5.53 as

$$R'(d) = \frac{2A(d)}{B(d)} + C$$

where

$$A(d) = \frac{1}{m} \sum_{j=1}^m I_j \lambda_j^{2d} \log(\lambda_j),$$

$$B(d) = \frac{1}{m} \sum_{j=1}^m I_j \lambda_j^{2d}$$

and

$$C = \frac{2}{m} \sum_{j=1}^m \log(\lambda_j).$$

Lemma 5.2.10 with $b = 1$ gives

$$A(d) = \left(-G_0 \left(\frac{2\pi m}{n} \right)^{2(d-d_0)} \right) \left(\frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma_0} \right) \quad (5.76)$$

$$+ O_p \left(\frac{1}{m^{1/2}} \max_{j \in (1, \dots, m)} \left| \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma_0} \right| \right)$$

and Lemma 5.2.10 with $b = 0$ gives

$$B(d) = \left(G_0 \left(\frac{2\pi m}{n} \right)^{2(d-d_0)} \right) \left(\frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0} \right. \quad (5.77)$$

$$\left. + O_p \left(\frac{1}{m^{1/2}} \max_{j \in (1, \dots, m)} \left| \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0} \right| \right) \right).$$

Note,

$$\frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0} = O \left(\max_{j \in (1, \dots, m)} \left| \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0} \right| \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} \right)$$

and use of standard growth identities gives

$$\frac{1}{m} \sum_{j=1}^m \left(\frac{j}{m} \right)^{2(d-d_0)} = O(1)$$

thus the first terms in each of equations 5.76 and 5.77 are of higher order as $m \rightarrow \infty$. Substitution of equations 5.76 and 5.77 back into $R'(d)$ thus gives the required result. \square

Lemma 5.2.12. *Under assumptions 5.2.1- 5.2.4, as $n \rightarrow \infty$, the first derivative with respect to d of $R(d)$ defined in equation 5.50 evaluated at the point d_0 is equal to*

$$R'(d_0) = -2\gamma + o_p(1) \text{ as } n \rightarrow \infty.$$

Proof. Lemma 5.2.11 gives

$$R'(d) = \frac{-2 \sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma}}{\sum_{j=1}^m \lambda_j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma}} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1).$$

and substitution of $d = d_0$ gives

$$R'(d_0) = \frac{-2 \frac{1}{m} \sum_{j=1}^m \log \left(\frac{1}{\lambda_j} \right)^{1+\gamma}}{\frac{1}{m} \sum_{j=1}^m \log \left(\frac{1}{\lambda_j} \right)^{\gamma}} - \frac{2}{m} \sum_{j=1}^m \log(\lambda_j) + o_p(1)$$

Phillips [2001] shows that for any slowly varying function $L(j)$,

$$\begin{aligned} \frac{1}{m} \sum_{j=1}^m L(j)^a &= L(m)^a + am^2 L''(m) L(m)^{a-1} + (a^2 - a)(mL'(m))^2 L(m)^{a-2} \\ &\quad + o\left((mL'(m))^2 L(m)^{a-2}\right) \end{aligned} \quad (5.78)$$

as $m \rightarrow \infty$. The use of this formula with $L(j) = \log\left(\frac{1}{\lambda_j}\right)$, $a = 1 + \gamma$, γ and $L(j) = \log(\lambda_j)$ and $a = 1$ gives the required result in a straightforward manner and is omitted to save space. \square

Lemma 5.2.13. *Under assumptions 5.2.1-5.2.4*

$$\log\left(\frac{\hat{G}(d, \gamma)}{G(d, \gamma)}\right) = O_p\left(\frac{1}{m^{1/2}}\right)$$

Proof. The inequality $|\log(1+x)| \leq 2|x|$ for $|x| \leq 0.5$ can be used to show

$$\log\left(\frac{\hat{G}(d, \gamma)}{G(d, \gamma)}\right) \leq 2 \left| \frac{\hat{G}(d, \gamma)}{G(d, \gamma)} - 1 \right| = 2 \left| \frac{A(d, \gamma)}{B(d, \gamma)} \right|$$

where

$$A(d, \gamma) = \frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)} \left(\frac{1_j}{g_j} - 1\right)$$

and

$$B(d, \gamma) = \frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)}$$

for $g_j = G_0 \lambda_j^{-2d_0} \log(1/\lambda_j)^{\gamma_0}$. Since

$$\frac{\log(1/\lambda_j)}{\log(1/\lambda_m)} = \frac{\log(n) - \log(2\pi j)}{\log(n) - \log(2\pi m)} \rightarrow 1, \quad \text{as } n \rightarrow \infty, \quad \forall j \in (1, \dots, m), \quad (5.79)$$

from assumption 5.2.4,

$$A(d, \gamma) \rightarrow A(d), \quad \text{as } n \rightarrow \infty$$

where

$$A(d) = \frac{2(d-d_0)+1}{m} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \left(\frac{l_j}{g_j} - 1\right),$$

and Lemma 5.2.5 shows this is $O(m^{-0.5})$ as $n \rightarrow \infty$. Consider $B(d, \gamma)$.

$$\inf_{\Theta_1} B(d, \gamma) \geq 1 - \sup_{\Theta_1} \left| \left(\frac{2(d-d_0)+1}{m \log\left(\frac{1}{\lambda_j}\right)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log\left(\frac{1}{\lambda_j}\right)^{(\gamma_0-\gamma)} \right) - 1 \right|.$$

Equation 5.79 can be used once again to show

$$\begin{aligned} \sup_{\Theta_1} \left| \left(\frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)} \right) - 1 \right| & \quad (5.80) \\ \rightarrow \sup_{\Theta_1} \left| \left(\frac{2(d-d_0)+1}{m} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \right) - 1 \right|, & \quad \text{as } n \rightarrow \infty \end{aligned}$$

and Robinson [1995b] Lemma 1 proves that equation 5.80 is $O(1/m^{-0.5})$. Therefore

$$\inf_{\Theta_1} B(d, \gamma) \geq 1 - o(1) > 0 \text{ as } n \rightarrow \infty,$$

implying

$$\left| \frac{A(d, \gamma)}{B(d, \gamma)} \right| = O_p\left(\frac{1}{m^{1/2}}\right), \text{ as } n \rightarrow \infty,$$

which completes the proof. \square

Lemma 5.2.14. *Under assumption 5.2.4*

$$\sup_{\Theta_1} \left(\log \left(\frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)} \right) \right) = O\left(\frac{1}{m^{1/2}}\right),$$

as $n \rightarrow \infty$.

Proof. The inequality $|\log(1+x)| \leq 2|x|$ for $|x| \leq 0.5$ can be used to show that

$$\begin{aligned} & \sup_{\Theta_1} \left(\log \left(\frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)} \right) \right) \\ & \leq \sup_{\Theta_1} \left| \left(\frac{2(d-d_0)+1}{m \log(1/\lambda_m)^{(\gamma_0-\gamma)}} \sum_{j=1}^m \left(\frac{j}{m}\right)^{2(d-d_0)} \log(1/\lambda_j)^{(\gamma_0-\gamma)} \right) - 1 \right| \end{aligned}$$

and under assumption 5.2.4, arguing as with equation 5.80, Robinson [1995b] Lemma 1 proves that this is $O(1/m^{-0.5})$. □

Lemma 5.2.15. *As $m \rightarrow \infty$,*

$$\left| \frac{1}{m} \sum_{j=1}^m \log(j) - (\log(m) - 1) \right| = O\left(\frac{1}{m^{1/2}}\right)$$

This is shown by Robinson [1995b] Lemma 2.

Lemma 5.2.16. *As $m \rightarrow \infty$,*

$$\left| \frac{1}{m} \sum_{j=1}^m \log(\log(1/\lambda_j)) - \log(\log(1/\lambda_m)) \right| = O\left(\frac{1}{\log(\lambda_m)}\right) \quad (5.81)$$

The proof follows directly using equation 5.78, see Phillips [2001].

Lemma 5.2.17. *Under assumption 5.2.4, as $n \rightarrow \infty$,*

$$\begin{aligned} \hat{D}(d, \gamma) &= \left(\frac{1}{m} \sum_{j=1}^m \log(1/\lambda_j)^{(\gamma-\gamma_0)} \right) \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p}\right)^{2(d-d_0)} j^{2d_0} \log(1/\lambda_j)^{-\gamma} I_j \\ &= (1 + o(1)) \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p}\right)^{2(d-d_0)} j^{2d_0} \log(1/\lambda_j)^{-\gamma} I_j \end{aligned}$$

Proof. From equation 5.78, see Phillips [2001],

$$\frac{1}{m} \sum_{j=1}^m \log(1/\lambda_j)^{(\gamma-\gamma_0)} = \log(1/\lambda_m)^{(\gamma-\gamma_0)} (1 + o(1))$$

From assumption 5.2.4, for all $j \in (1, \dots, m)$,

$$\log(1/\lambda_m)^{(\gamma-\gamma_0)} \log(1/\lambda_j)^{-\gamma} = \log(1/\lambda_m)^{(\gamma-\gamma_0-\gamma)} + o(1) = \log(1/\lambda_m)^{-\gamma_0} + o(1)$$

Finally, from assumption 5.2.4, for all $j \in (1, \dots, m)$,

$$\log(1/\lambda_m)^{-\gamma_0} = \log(1/\lambda_j)^{-\gamma_0} + o(1)$$

thus

$$\hat{D}(d, \gamma) = (1 + o(1)) \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p}\right)^{2(d-d_0)} j^{2d_0} I_j \log(1/\lambda_j)^{-\gamma_0},$$

as required. □

Lemma 5.2.18. For

$$p = p_m = \exp(m^{-1} \sum_{j=1}^m \log(j)) \quad (5.82)$$

and

$$\tilde{D}(d) = \frac{1}{m} \sum_{j=1}^m \left(\frac{j}{p}\right)^{2(d-d_0)} j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j,$$

then, for $d \in \Theta_2$ defined in the proof of Theorem 5.2.2 and $d_0 > -\nu$, as $n \rightarrow \infty$,

$$P\left(\frac{\tilde{D}(d)}{\tilde{D}(d_0)} - 1 < 0\right) \leq P\left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j < 0\right),$$

where

$$a_j = \begin{cases} \left(\frac{j}{p}\right)^{(2\eta-1)}, & 1 \leq j \leq p \\ \left(\frac{j}{p}\right)^{2(-d_0-0.5)}, & p < j \leq m \end{cases} \quad (5.83)$$

Proof. Note,

$$\tilde{D}(d_0) = \frac{1}{m} \sum_{j=1}^m j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j > 0$$

since $j, I_j > 0$ and $\lambda_j < 1$ as $n \rightarrow \infty$. Thus

$$P\left(\frac{\tilde{D}(d)}{\tilde{D}(d_0)} - 1 < 0\right) = P\left(\tilde{D}(d) - \tilde{D}(d_0) < 0\right).$$

From the definition of p , $1 \leq p \leq m$, and for $d \in \Theta_2$ and $d_0 > -\nu$

$$-d_0 - 0.5 \geq d - d_0 \geq \nu - 0.5.$$

Since $\inf_{\Theta_2} (j/p)^{2(d-d_0)} \geq (j/p)^{2(\eta-1)}$ for $1 \leq j \leq p$, while $\inf_{\Theta_2} (j/p)^{2(d-d_0)} > (j/p)^{2(-d_0-0.5)}$ for $p < j \leq m$, it follows that

$$\inf_{\Theta_2} \tilde{D}(d) > \frac{1}{m} \sum_{j=1}^m a_j j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j,$$

which implies

$$\inf_{\Theta_2} (\tilde{D}(d) - \tilde{D}(d_0)) > \frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j.$$

Thus

$$P\left(\tilde{D}(d) - \tilde{D}(d_0) < 0\right) \leq P\left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j < 0\right),$$

as required. \square

Lemma 5.2.19. For $d \in \Theta_2$ defined in the proof of Theorem 5.2.2, $d_0 > -\nu$ and p, a_j and g_j defined in equations 5.82, 5.83 and 5.71 respectively, as $m \rightarrow \infty$,

$$P\left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j < 0\right) \leq P\left(\left|\frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1\right)\right| \geq 1\right)$$

Proof. Note, since $2\pi/n > 0$ and $\lambda_j = 2\pi j/n$

$$P\left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j < 0\right) = P\left(\frac{1}{m} \sum_{j=1}^m (a_j - 1) \frac{I_j}{g_j} < 0\right)$$

$$\leq P \left(\left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1 \right) \right| \geq \left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \right| \right)$$

From Lemma 5.2.15 $p \sim \exp(\log(m) - 1) = m/e$ as $m \rightarrow \infty$ giving

$$\sum_{j=1}^p a_j \sim p^{1-2\eta} \int_0^p x^{2\eta-1} dx \sim \frac{m}{2e\eta} \quad \text{as } n \rightarrow \infty.$$

Hence

$$\frac{1}{m} \sum_{j=1}^m a_j - 1 \geq \frac{1}{m} \sum_{j=1}^p a_j - 1 \sim \frac{1}{2e\eta} - 1 \quad \text{as } n \rightarrow \infty. \quad (5.84)$$

and η can be chosen such that equation 5.84 is greater than 1. Therefore

$$P \left(\left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1 \right) \right| \geq \left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \right| \right) \leq P \left(\left| \frac{1}{m} \sum_{j=1}^m (a_j - 1) \left(\frac{I_j}{g_j} - 1 \right) \right| \geq 1 \right),$$

which completes the proof. \square

The following definitions will be used during the following Lemmas.

$$G_{ik}(d, \gamma) = \frac{2^i(-1)^k}{m} \sum_{j=1}^m \lambda_j^{2d} \log(1/\lambda_j)^{-\gamma} \log(\lambda_j)^i \log(\log(1/\lambda_j))^k I_j,$$

$$F_{ik}(d, \gamma) = \frac{2^i(-1)^k}{m} \sum_{j=1}^m \lambda_j^{2d} \log(1/\lambda_j)^{-\gamma} \log(j)^i \log(\log(1/\lambda_j))^k I_j$$

and

$$E_{ik}(d, \gamma) = \frac{2^i(-1)^k}{m} \sum_{j=1}^m j^{2d} \log(1/\lambda_j)^{-\gamma} \log(j)^i \log(\log(1/\lambda_j))^k I_j. \quad (5.85)$$

Note,

$$E_{ik}(d, \gamma) = \left(\frac{n}{2\pi} \right)^{2d} F_{ik}(d, \gamma) \quad (5.86)$$

and, for $i = 0$,

$$E_{0k}(d, \gamma) = \left(\frac{n}{2\pi}\right)^{2d} E'_{0k}(d, \gamma) = \left(\frac{n}{2\pi}\right)^{2d} G_{0k}(d, \gamma) \quad (5.87)$$

Lemma 5.2.20. *Under assumptions 5.2.5-5.2.8, for a constant ε , such that $0 < 2\varepsilon < \log(m)^2$ and the sets $M = ((d, \gamma) : (|d - d_0|, |\gamma - \gamma_0|) \leq \varepsilon \log(m)^{(-3-\nu)}(1, 1))$, for ν defined in assumption 5.2.8 and $\bar{M} = \mathbb{R}^2 - M$,*

$$\begin{aligned} & P \left(\left| E_{ik}(\tilde{d}, \tilde{\gamma}) - E_{ik}(d_0, \gamma_0) \right| > \eta \left(\frac{n}{2\pi}\right)^{2d_0} \right) \\ & \leq P \left(G_{00}(d_0, \gamma_0) > \left| \frac{\eta \log(2\pi/n)^{2+\nu-i}}{e\varepsilon \log(\log(n/2\pi))^k} + \frac{2\eta \log(2\pi/n)^{3+\nu-i}}{e\varepsilon \log(\log(n/2\pi))^{k+1}} \right| \right) + P \left((\tilde{d}, \tilde{\gamma}) \in \bar{M} \right) \\ & \rightarrow 0, \text{ as } n \rightarrow \infty \end{aligned}$$

Proof. Note, using the identity

$$x \leq (1+x) \log(1+x),$$

for $x \geq 0$, gives

$$\left(j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0-\gamma} - 1 \right) \leq j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0-\gamma} \log \left(j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0-\gamma} \right).$$

For $(d, \gamma) \in M$, $|d - d_0|, |\gamma - \gamma_0| \leq \varepsilon \log(m)^{(-3-\nu)}$ by definition, thus, for $\lambda_j = 2\pi j/n$ and $j \in (1, \dots, m)$

$$j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0-\gamma} \leq \left(m \log \left(\frac{n}{2\pi} \right) \right)^{2\varepsilon \log(m)^{(-3-\nu)}},$$

and, under assumption 5.2.8,

$$\begin{aligned} & \log \left(\left(m \log \left(\frac{n}{2\pi} \right) \right)^{2\varepsilon \log(m)^{(-3-\nu)}} \right) = 2\varepsilon \log(m)^{(-3-\nu)} \log \left(m \log \left(\frac{n}{2\pi} \right) \right) \\ & = 2\varepsilon \log(m)^{(-3-\nu)} \left(\log(m) + \log \left(\log \left(\frac{n}{2\pi} \right) \right) \right) = o(1). \end{aligned}$$

Hence, for some constant $a > 1$,

$$\left(j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0 - \gamma} - 1 \right) \leq a \log \left(j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0 - \gamma} \right). \quad (5.88)$$

Use of equation 5.88 and the definition of $E_{ik}(d, \gamma)$ given in equation 5.85 gives, for $(d, \gamma) \in M$,

$$\begin{aligned} & |E_{ik}(d, \gamma) - E_{ik}(d_0, \gamma_0)| \\ & \leq a \left| \frac{2^i}{m} \sum_{j=1}^m \log \left(j^{2(d-d_0)} \log \left(\frac{1}{\lambda_j} \right)^{\gamma_0 - \gamma} \right) \log(j)^i \log \left(\log \left(\frac{1}{\lambda_j} \right) \right)^k j^{2d_0} \log \left(\frac{1}{\lambda_j} \right)^{-\gamma_0} I_j \right| \\ & = a |d - d_0| |E_{(i+1),k}(d_0, \gamma_0)| + \frac{a}{2} |\gamma - \gamma_0| |E_{i,(k+1)}(d_0, \gamma_0)|. \end{aligned}$$

Note, for $i, k \geq 0$,

$$|E_{(i+1),k}(d_0, \gamma_0)| \leq \left| \log(m)^i \log(\log(n/2\pi))^k \right| |E_{00}(d_0, \gamma_0)|.$$

Also, for $(d, \gamma) \in M$, $|d - d_0|, |\gamma - \gamma_0| \leq \varepsilon \log(m)^{(-3-\nu)}$ by definition, thus,

$$\begin{aligned} & |E_{ik}(d, \gamma) - E_{ik}(d_0, \gamma_0)| \\ & \leq a\varepsilon \left| \log(m)^{i-2-\nu} \log(\log(n/2\pi))^k + 2 \log(m)^{i-3-\nu} \log(\log(n/2\pi))^{k+1} \right| |E_{00}(d_0, \gamma_0)| \\ & = a\varepsilon \left| \log(m)^{i-2-\nu} \log(\log(n/2\pi))^k + 2 \log(m)^{i-3-\nu} \log(\log(n/2\pi))^{k+1} \right| \left(\frac{n}{2\pi} \right)^{2d_0} G_{00}(d_0, \gamma_0), \end{aligned}$$

where the last equality is from the property given in equation 5.87. Thus, as $n \rightarrow \infty$, for $\bar{M} = \mathbb{R}^2 - M$ and all $\eta > 0$,

$$\begin{aligned} & P \left(\left| E_{ik}(\tilde{d}, \tilde{\gamma}) - E_{ik}(d_0, \gamma_0) \right| > \eta \left(\frac{n}{2\pi} \right)^{2d_0} \right) \quad (5.89) \\ & \leq P \left(G_{00}(d_0, \gamma_0) > \left| \frac{\eta \log(2\pi/n)^{2+\nu-i}}{a\varepsilon \log(\log(n/2\pi))^k} + \frac{2\eta \log(2\pi/n)^{3+\nu-i}}{a\varepsilon \log(\log(n/2\pi))^{k+1}} \right| \right) + P \left((\tilde{d}, \tilde{\gamma}) \in \bar{M} \right). \end{aligned}$$

Note, for $i, k \in (0, 1, 2)$,

$$\left| \frac{\eta \log(2\pi/n)^{2+\nu-i}}{a\varepsilon \log(\log(n/2\pi))^k} + \frac{2\eta \log(2\pi/n)^{3+\nu-i}}{a\varepsilon \log(\log(n/2\pi))^{k+1}} \right| \rightarrow \infty, \text{ as } n \rightarrow \infty$$

and substitution of $d = d_0$ and $\gamma = \gamma_0$ into Lemma 5.2.13 shows that $G_{00}(d_0, \gamma_0) \rightarrow G_0 < \infty$ as $n \rightarrow \infty$, thus the first probability on the right hand side of equation 5.89 tends to zero. The second probability tends to zero from the arguments of Theorem 5.2.2. □

Lemma 5.2.21. *Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$\begin{aligned} & \left| I'_{ik}(d_0, \gamma_0) - G_0 \frac{2^i(-1)^k}{m} \sum_{j=1}^m \log(j)^i \log(\log(1/\lambda_j))^k \right| \\ &= O_p \left(\frac{\log(m)^2 \log \log(1/\lambda_m)^2}{m^{1/2}} \right) = o_p(1), \text{ for } i, k \in (0, 1, 2) \end{aligned}$$

Proof. Note for $i, k \in (0, 1, 2)$, $r \in (1, \dots, m-1)$,

$$\left| \log(r)^i \log(\log(1/\lambda_r))^k - \log(r+1)^i \log(\log(1/\lambda_{r+1}))^k \right| = O(\log(m)^2 \log(\log(1/\lambda_m))^2). \quad (5.90)$$

By definition,

$$\begin{aligned} & \left| I'_{ik}(d_0, \gamma_0) - G_0 \frac{2^i(-1)^k}{m} \sum_{j=1}^m \log(j)^i \log \left(\log \left(\frac{1}{\lambda_j} \right) \right)^k \right| \\ &= \frac{G_0 2^i}{m} \sum_{j=1}^m \left| \log(j)^i \log \left(\log \left(\frac{1}{\lambda_j} \right) \right)^k \left(\frac{I_j}{g_j} - 1 \right) \right|, \end{aligned}$$

which, by summation by parts is,

$$\begin{aligned} & \leq \frac{G_0 2^i}{m} \sum_{r=1}^{m-1} \left| \log(r)^i \log(\log(1/\lambda_r))^k - \log(r+1)^i \log(\log(1/\lambda_{r+1}))^k \right| \left| \sum_{j=1}^r \left(\frac{I_j}{g_j} - 1 \right) \right| \\ & \quad + \frac{G_0 2^i}{m} \log(m)^i \log \log(1/\lambda_m)^k \left| \sum_{j=1}^m \left(\frac{I_j}{g_j} - 1 \right) \right| \end{aligned}$$

and use of Lemma 5.2.6 and substitution of equation 5.90 show this is

$$= O_p \left(\frac{\log(m)^2 \log \log(1/\lambda_m)^2}{m^{1/2}} \right) \text{ for } i, k \in (0, 1, 2), \text{ as } n \rightarrow \infty.$$

Under assumption 5.2.8 this is $o_p(1)$ as required. \square

Lemma 5.2.22. Let $R_{dd}(d, \gamma)$ be the second partial derivative of $R_2(d, \gamma)$ with respect to d twice. Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,

$$R_{dd}(\tilde{d}, \tilde{\gamma}) = 4 + o_p \left(\frac{1}{\log(m)} \right),$$

Proof. From the definition of $E_{ik}(d, \gamma)$ given in equation 5.85, $R_{dd}(\tilde{d}, \tilde{\gamma})$ is given by

$$R_{dd}(\tilde{d}, \tilde{\gamma}) = \frac{E_{2,0}(\tilde{d}, \tilde{\gamma})E_{0,0}(\tilde{d}, \tilde{\gamma}) - E_{1,0}^2(\tilde{d}, \tilde{\gamma})}{E_{0,0}^2(\tilde{d}, \tilde{\gamma})}.$$

Thus, use of Lemma 5.2.20 gives

$$R_{dd}(\tilde{d}, \tilde{\gamma}) = \frac{(E_{2,0}(d_0, \gamma_0) + o_p(n^{2\tilde{d}}))(E_{0,0}(d_0, \gamma_0) + o_p(n^{2\tilde{d}})) - (E_{1,0}(d_0, \gamma_0) + o_p(n^{2\tilde{d}}))^2}{(E_{0,0}(d_0, \gamma_0) + o_p(n^{2\tilde{d}}))^2}$$

dividing numerator and denominator by $(n/2\pi)^{4\tilde{d}}$ and using the property given in equation 5.86 gives this is

$$= \frac{E_{2,0}^1(d_0, \gamma_0)E_{0,0}^1(d_0, \gamma_0) - E_{1,0}^2(d_0, \gamma_0)}{E_{0,0}^2(d_0, \gamma_0)} + o_p(1)$$

and use of Lemma 5.2.21 gives this is

$$= 4 \left(\frac{1}{m} \sum_1^m \log(j)^2 - \left(\frac{1}{m} \sum_1^m \log(j) \right)^2 \right) (1 + o_p(1)) \text{ as } n \rightarrow \infty.$$

Using the extended version of equation 5.78 given in the proof of Phillips [2001] Lemma 7.3 shows that for any slowly varying function $L(j)$,

$$\frac{1}{m} \sum_{j=1}^m L(j)^2 - \left(\frac{1}{m} \sum_{j=1}^m L(j) \right)^2 = -3(mL'(m))^2 - 4m^3 L'(m)L''(m) + o \left(\frac{(mL'(m))^3}{L(m)} \right) \quad (5.91)$$

and taking $L(j) = \log(j)$ gives $mL'(m) = 1$ and $m^3L'(m)L''(m) = -1$, therefore

$$R_{dd}(\bar{d}, \tilde{\gamma}) = 4 \left(1 + o\left(\frac{1}{\log(m)}\right) \right) (1 + o_p(1)) = 4 + o_p\left(\frac{1}{\log(m)}\right), \text{ as } n \rightarrow \infty,$$

as required. \square

The proofs of Lemmas 5.2.23 and 5.2.24 are similar to that of Lemma 5.2.22 and thus omitted.

Lemma 5.2.23. *Let $R_{d\gamma}(d, \gamma)$ be the second partial derivative of $R_2(d, \gamma)$ with respect to d then γ . Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$R_{d\gamma}(\bar{d}, \tilde{\gamma}) = \frac{-2}{\log(2\pi m/n)} - \frac{4}{\log(2\pi m/n)^2} + o_p\left(\frac{1}{\log(2\pi m/n)^2} \log(\log(n/(2\pi m)))\right).$$

Lemma 5.2.24. *Let $R_{\gamma\gamma}(d, \gamma)$ be the second partial derivative of $R_2(d, \gamma)$ with respect to γ twice. Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$R_{\gamma\gamma}(\bar{d}, \tilde{\gamma}) = \frac{1}{\log(2\pi m/n)^2} + \frac{4}{\log(2\pi m/n)^3} + o_p\left(\frac{1}{\log(2\pi m/n)^3 \log(\log(n/(2\pi m)))}\right)$$

Lemma 5.2.25. *Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$\frac{m^{1/2}}{2} R_d(d_0, \gamma_0) \rightarrow_d N(0, 1)$$

Proof. Standard calculations give

$$\begin{aligned} \frac{m^{1/2}}{2} R_d(d_0, \gamma_0) &= \frac{1}{m^{1/2}} \left(\frac{\sum_{j=1}^m \log(\lambda_j) \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j}{1/m \sum_{j=1}^m \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j} - \sum_{j=1}^m \log(\lambda_j) \right) \\ &= \frac{1}{m^{1/2}} \left(\frac{\sum_{j=1}^m \left(\log(\lambda_j) \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j - (1/m \sum_{s=1}^m \log(\lambda_s)) \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j \right)}{1/m \sum_{j=1}^m \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j} \right) \end{aligned}$$

and substitution of $d = d_0$ and $\gamma = \gamma_0$ into Lemma 5.2.13 shows that $G_{00}(d_0, \gamma_0) \rightarrow G_0$, thus this is

$$= \frac{1}{m^{1/2}} \left(\frac{\sum_{j=1}^m \left(\log(\lambda_j) \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j - (1/m \sum_{s=1}^m \log(\lambda_s)) \lambda_j^{2d_0} \log(1/\lambda_j)^{-\gamma_0} I_j \right)}{G_0 + o(1)} \right)$$

Letting $v_j = \log(j) - 1/m \sum_{s=1}^m \log(s)$, with $\sum_{j=1}^m v_j = 0$, this is

$$\begin{aligned} &= \frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \frac{I_j}{g_j} \right) (1 + o(1)). \quad (5.92) \\ &= \frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \left(\frac{I_j}{g_j} - 1 \right) \right) (1 + o(1)). \end{aligned}$$

Note, v_j is defined exactly as in the proof of Robinson [1995b] and thus Lemma 5.2.7 is applicable giving,

$$\frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \left(\frac{I_j}{g_j} - 1 \right) \right) (1 + o(1)) \rightarrow_d N(0, 1)$$

as required. This completes the proof. □

Lemma 5.2.26. *Under assumptions 5.2.5-5.2.8, for*

$$\eta_j = -\log(2\pi m/n) \left(\log(\log(1/\lambda_j)) - 1/m \sum_{s=1}^m \log(\log(1/\lambda_s)) \right),$$

then, as $n \rightarrow \infty$,

$$\frac{1}{m^{1/2}n} \sum_{j=1}^m |\eta_j| = o(1) \quad (5.93)$$

and

$$\frac{1}{m} \sum_{j=1}^m \eta_j^2 = 1 + o(1). \quad (5.94)$$

Proof. Note, from the definition of η_j

$$\begin{aligned} \frac{1}{m^{1/2}n} \sum_{j=1}^m |\eta_j| &= \left| \frac{\log(2\pi m/n)m^{1/2}}{n} \right| \left| \frac{1}{m} \sum_{j=1}^m \left| \log(\log(1/\lambda_j)) - \frac{1}{m} \sum_{s=1}^m \log(\log(1/\lambda_s)) \right| \right| \\ &\leq 2 \left| \frac{\log(2\pi m/n)m^{1/2}}{n} \right| \left| \frac{1}{m} \sum_{j=1}^m |\log(\log(1/\lambda_j))| \right| \\ &= O\left(\frac{\log(2\pi m/n)m^{1/2} \log(\log(1/\lambda_m))}{n}\right) \end{aligned}$$

thus, from assumption 5.2.8, equation 5.93 is satisfied.

Next

$$\begin{aligned} \frac{1}{m} \sum_{j=1}^m \eta_j^2 &= \log(2\pi m/n)^2 \left(\frac{1}{m} \sum_{j=1}^m \log(\log(1/\lambda_j))^2 - \left(\frac{1}{m} \sum_{s=1}^m \log(\log(1/\lambda_s)) \right)^2 \right), \\ &= \log(2\pi m/n)^2 \left(\frac{1}{\log(2\pi m/n)^2} + o\left(\frac{1}{\log(2\pi m/n)^3}\right) \right) = 1 + o(1) \end{aligned}$$

thus equation 5.94 is satisfied

□

The proof of Lemma 5.2.27 is similar to Lemma 5.2.25 with the replacement of v_j by η_j made possible due to Lemma 5.2.26 and is thus omitted.

Lemma 5.2.27. *Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$\log(2\pi m/n)m^{1/2}R_\gamma(d_0, \gamma_0) \rightarrow_d N(0, 1)$$

Lemma 5.2.28. *Under assumptions 5.2.5-5.2.8, for*

$$v_j = \log(j) - \frac{1}{m} \sum_{s=1}^m \log(s)$$

and

$$\eta_j = -\log(2\pi m/n) \left(\log(\log(1/\lambda_j)) - \frac{1}{m} \sum_{s=1}^m \log(\log(1/\lambda_s)) \right),$$

then, as $n \rightarrow \infty$,

$$\frac{-1}{m} \sum_{j=1}^m v_j \eta_j = 1 + o(1),$$

Proof.

$$\begin{aligned} \frac{-1}{m} \sum_{j=1}^m v_j \eta_j &= \log(2\pi m/n) \left(\frac{1}{m} \sum_1^m \log(j) \log(\log(1/\lambda_j)) - \frac{1}{m} \sum_1^m \log(j) \frac{1}{m} \sum_1^m \log(\log(1/\lambda_j)) \right), \\ &= \log(2\pi m/n) \left(\frac{1}{\log(2\pi m/n)} + o\left(\frac{1}{\log(2\pi m/n)^2}\right) \right) = 1 + o(1), \end{aligned}$$

as required. This completes the proof. \square

Lemma 5.2.29. *Under assumptions 5.2.5-5.2.8, as $n \rightarrow \infty$,*

$$Acov\left(m^{1/2}R_d(d_0, \gamma_0), \log(2\pi m/n)m^{1/2}R_\gamma(d_0, \gamma_0)\right) \rightarrow -1$$

where $Acov(A, B)$ is the asymptotic covariance of A and B .

Proof. Let $A = m^{1/2}R_d(d_0, \gamma_0)$ and $B = \log(2\pi m/n)m^{1/2}R_\gamma(d_0, \gamma_0)$. The proofs of Lemmas 5.2.25 and 5.2.27 show, as $n \rightarrow \infty$,

$$A \rightarrow X = \frac{1}{m^{1/2}} \left(\sum_{j=1}^m v_j \left(\frac{l_j}{g_j} - 1 \right) \right)$$

and

$$B \rightarrow Y = \frac{1}{m^{1/2}} \left(\sum_{j=1}^m \eta_j \left(\frac{l_j}{g_j} - 1 \right) \right)$$

and $E(X) \rightarrow 0$ and $E(Y) \rightarrow 0$. Lemma 5.2.28 shows Lemma 5.2.8 is applicable and thus $E(XY) \rightarrow -1$, as $n \rightarrow \infty$. Therefore, as $n \rightarrow \infty$, $cov(X, Y) \rightarrow -1$ and thus $Acov(A, B) \rightarrow -1$ as required. \square

5.3 Dual Parameter FARIMA model

The discussion of the GPII and Local Whittle estimates in sections 5.1 and 5.2 show the importance of considering the dual parameter long memory case when $\gamma \neq 0$. Under the assumptions given in sections 5.1 and 5.2, Theorems 5.1.3 and 5.2.1 show the standard GPII and stand Local Whittle methods of estimation, respectively, produce estimates with biases which reduce at a slower asymptotic rate than their variances as the series length $n \rightarrow \infty$, whilst Theorems 5.1.4 and 5.2.2 show the new dual parameter GPII and dual parameter Local Whittle methods, respectively, produce asymptotically consistent estimates of both d and γ . However, Theorems 5.1.4 and 5.2.3 show the asymptotic distributions to be collinear and the variances to decay at a slower asymptotic rate, making these estimates only suitable for longer time series.

In this section a parametric approach to estimating the dual long memory parameters is introduced, based on the new dual parameter FARIMA (DFARIMA) model, an extension of the FARIMA model which allows for the spectrum near $\lambda = 0$ to be of the form given in equation 5.61. Define now a DFARIMA(p, d, c, q) process, y_t , as

$$y_t = (1 - L)^{-d} (1 - \log(1 - L))^{-c} \frac{\theta(L)}{\phi(L)} \varepsilon_t \quad (5.95)$$

where L is the lag operator, $\theta(L)$, $\phi(L)$ are polynomials of order p and q respectively with no common roots, such that $\theta(z) \neq 0$, $\phi(z) \neq 0$ for $|z| < 1$ and ε_t is a white noise sequence with finite variance.

This is a linear model, which can include the boundary conditions of long memory at $d = 0$ and $d = 0.5$ and reduces to a FARIMA model if the new parameter, c , is set to zero. The spectral density of this process as $\lambda \rightarrow 0$ is given in Theorem 5.3.1, the asymptotic infinite AR and MA representations are given in Theorem 5.3.2 and the asymptotic autocorrelations are given in Theorem 5.3.3. Proposition 5.3.1 then shows that this new model still fits the criteria of Fox and Taqqu [1986], Dahlhaus [1989] and Giraitis and Surgailis [1990] and thus that the asymptotic sampling distributions given in these papers for the exact and Whittle likelihood methods still hold, that is, that the estimates of d and the new parameter c are \sqrt{n} consistent and efficient in the sense of Fisher. Numerical values are given for the asymptotic sampling distribution of the estimates of d and c from DFARIMA(0, $d, c, 0$) process and this is compared with the sampling distribution of d from a DFARIMA(0, $d, 0, 0$) process as an example.

The following theorem now gives properties of the spectral density near $\lambda = 0$.

Theorem 5.3.1. *Let y_t be a DFARIMA(p, d, c, q) process defined by equation 5.95. Then the spectrum, $f_y(\lambda)$, of y_t satisfies*

$$f_y(\lambda) \sim G\lambda^{-2d} \log\left(\frac{1}{\lambda}\right)^{-2c} \quad \text{as } \lambda \rightarrow 0.$$

where $0 < G < \infty$ is a constant as $\lambda \rightarrow 0$.

Proof. From equation 5.95 the spectrum, $f_y(\lambda)$, is given by

$$f_y(\lambda) = |1 - e^{-i\lambda}|^{-2d} \left| \log \frac{e}{1 - e^{-i\lambda}} \right|^{-2c} \frac{|\theta(e^{-i\lambda})|^2 \sigma^2}{|\phi(e^{-i\lambda})|^2 2\pi} \quad (5.96)$$

Standard results give

$$|1 - e^{-i\lambda}|^2 = (1 - e^{-i\lambda})(1 - e^{i\lambda}) = (2 \sin(\lambda/2))^2 \rightarrow \lambda^2 \text{ as } \lambda \rightarrow 0 \quad (5.97)$$

and

$$\frac{|\theta(e^{-i\lambda})|^2 \sigma^2}{|\phi(e^{-i\lambda})|^2 2\pi} \rightarrow G \text{ as } \lambda \rightarrow 0 \quad (5.98)$$

where $0 < G < \infty$ is a constant as $\lambda \rightarrow 0$.

Also, converting to polar coordinates gives

$$1 - e^{-i\lambda} = 2 \sin(\lambda/2) e^{-i \tan^{-1}\left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)}\right)},$$

$$1 - e^{i\lambda} = 2 \sin(\lambda/2) e^{i \tan^{-1}\left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)}\right)},$$

where \tan^{-1} is the inverse tan function. Use of these gives

$$\begin{aligned} \left| \log \frac{e}{1 - e^{-i\lambda}} \right|^2 &= (1 - \log(1 - e^{-i\lambda}))(1 - \log(1 - e^{i\lambda})) \quad (5.99) \\ &= 1 - \log(1 - e^{-i\lambda}) - \log(1 - e^{i\lambda}) + \log(1 - e^{-i\lambda}) \log(1 - e^{i\lambda}) \\ &= \log\left(\frac{e}{(1 - e^{-i\lambda})(1 - e^{i\lambda})}\right) + \log(1 - e^{-i\lambda}) \log(1 - e^{i\lambda}) \\ &= \log\left(\frac{e}{(2 \sin(\lambda/2))^2}\right) + \log\left(2 \sin(\lambda/2) e^{-i \tan^{-1}\left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)}\right)}\right) \log\left(2 \sin(\lambda/2) e^{i \tan^{-1}\left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)}\right)}\right) \\ &= \log\left(\frac{e}{\left(2 \sin\left(\frac{\lambda}{2}\right)\right)^2}\right) \end{aligned}$$

$$\begin{aligned}
& + \left(\log \left(2 \sin \left(\frac{\lambda}{2} \right) \right) - i \tan^{-1} \left(\frac{\cos \left(\frac{\lambda}{2} \right)}{\sin \left(\frac{\lambda}{2} \right)} \right) \right) \left(\log \left(2 \sin \left(\frac{\lambda}{2} \right) \right) + i \tan^{-1} \left(\frac{\cos \left(\frac{\lambda}{2} \right)}{\sin \left(\frac{\lambda}{2} \right)} \right) \right) \\
& = \log \left(2 \sin(\lambda/2) \right)^2 + \log \left(\frac{e}{4 \sin(\lambda/2)^2} \right) + \left(\tan^{-1} \left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)} \right) \right)^2 \\
& = \left(-\log \left(\frac{1}{2 \sin(\lambda/2)} \right) \right)^2 + \log \left(\frac{e}{4 \sin(\lambda/2)^2} \right) + \left(\tan^{-1} \left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)} \right) \right)^2 \\
& = \log \left(\frac{1}{2 \sin(\lambda/2)} \right)^2 + \log \left(\frac{e}{4 \sin(\lambda/2)^2} \right) + \left(\tan^{-1} \left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)} \right) \right)^2 \\
& \dots \rightarrow \log \left(\frac{1}{\lambda} \right)^2 \quad \text{as } \lambda \rightarrow 0.
\end{aligned}$$

Substitution of the limits given in equations 5.97, 5.98 and 5.99 into equation 5.96 then gives

$$f_y(\lambda) \sim G\lambda^{-2d} \log \left(\frac{1}{\lambda} \right)^{-2c} \quad \text{as } \lambda \rightarrow 0.$$

as required. \square

Note, comparison of Theorem 5.3.1 to equation 5.61 shows γ is equivalent to $-2c$. Dual parameter GPII and dual parameter Local Whittle estimates of c , \hat{c}_{GPH} and \hat{c}_{LW} , respectively, can be found using $\hat{c}_{GPH} = -\hat{\gamma}_{GPH}/2$ and $\hat{c}_{LW} = -\hat{\gamma}_{LW}/2$, where $\hat{\gamma}_{GPH}$ and $\hat{\gamma}_{LW}$ are the dual parameter GPII and dual parameter Local Whittle estimates of γ given in sections 5.1 and 5.2, respectively. From this point on the c notation is adopted.

Theorem 5.3.2. *Let y_t be a DFARIMA(p, d, c, q) process as defined by equation 5.95. Then*

$$y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$$

where $\theta_0 = 1$ and, for $d \in [-0.5, 0) \cup (0, 0.5]$,

$$\theta_j \sim B \frac{j^{d-1}}{\Gamma(d)} \log(j)^{-c}, \quad \text{as } j \rightarrow \infty$$

while, for $d = 0, c \neq 0$

$$\theta_j \sim B j^{-1} \log(j)^{-c-1}, \quad \text{as } j \rightarrow \infty$$

and

$$y_t = \sum_{j=1}^{\infty} \phi_j y_{t-j} + \varepsilon_t$$

where for $d \in [-0.5, 0) \cup (0, 0.5]$,

$$\phi_j \sim B \frac{j^{-d-1}}{\Gamma(-d)} \log(j)^c, \quad \text{as } j \rightarrow \infty$$

while, for $d = 0, c \neq 0$

$$\phi_j \sim B j^{-1} \log(j)^{c-1}, \quad \text{as } j \rightarrow \infty$$

for some B such that $0 < B < \infty$.

The proof of this follows directly from Zygmund [1988], Theorem 2.31, page 192, and is thus omitted.

Theorem 5.3.3. Let y_t be a DFARIMA(p, d, c, q) process as defined by equation 5.95. Then,

a) for $d \in (0, 0.5)$

$$\gamma(u) \sim B u^{2d-1} \log(u)^{-2c}, \quad \text{as } u \rightarrow \infty$$

b) for $d = 0.5, c > 0.5$

$$\gamma(u) \sim B \log(u)^{1-2c}, \quad \text{as } u \rightarrow \infty$$

c) for $d = 0, c \neq 0$

$$\gamma(u) \sim B u^{-1} \log(u)^{-2c-1}, \quad \text{as } u \rightarrow \infty$$

for some $0 < B < \infty$, where $\gamma(u)$ is the autocorrelation function of y_t .

Proof. The proof of part a) follows directly from Theorem 5.3.2 and Proposition 4.3 of Inoue [1997] and is thus omitted.

b) Proposition 4.3 of Inoue [1997] shows when $\theta_u \sim u^{-p} l(u)$ as $u \rightarrow \infty$, for some slowly varying function $l(u)$ and constant $0.5 < p < 1$, then

$$\frac{\gamma(u)}{u \theta_u^2} \sim \int_0^{\infty} \frac{1}{(x^2 + x)^p} dx \quad \text{as } u \rightarrow \infty. \quad (5.100)$$

However, the bounds on p are not required to derive this integral, only to equate it to $B(2p-1, 1-p)$ where $B(p, q)$ is the Beta function, which cannot take arguments less than or equal to zero. Indeed, the proof is based on Bingham, Goldie and Teugels [1987] Corollary 1.4.2 and Theorem 1.5.2 both of which require only $p \in \mathbb{R}$, thus equation 5.100 still holds at the boundary conditions of $p = 0.5$ and $p = 1$ under the proof given by Inoue [1997].

From Theorem 5.3.2, when $d = 0.5$, $p = 0.5$. The integral is separated into two parts

$$\int_0^\infty \frac{1}{(x^2+x)^{0.5}} dx = \int_0^u \frac{1}{(x^2+x)^{0.5}} dx + \int_u^\infty \frac{1}{(x^2+x)^{0.5}} dx.$$

The second integral tends to zero as $u \rightarrow \infty$, leaving

$$\int_0^u \frac{1}{(x^2+x)^{0.5}} dx = \log(2\sqrt{u^2+u} + 2u + 1) \sim \log(u) \text{ as } u \rightarrow \infty.$$

Hence

$$\gamma(u) \sim u\theta_u^2 \log(u) \text{ as } u \rightarrow \infty,$$

and substitution of θ_u from Theorem 5.3.2 gives the required result.

c) Beginning again from equation 5.100, Theorem 5.3.2 shows $p = 1$ when $d = 0$. The integral is again split into two parts,

$$\int_0^\infty \frac{1}{(x^2+x)} dx = \int_0^{1/u} \frac{1}{(x^2+x)} dx + \int_{1/u}^\infty \frac{1}{(x^2+x)} dx,$$

where the first integral tends to zero as $u \rightarrow \infty$, leaving

$$\int_{1/u}^\infty \frac{1}{(x^2+x)} dx = \log(u+1) \sim \log(u) \text{ as } u \rightarrow \infty.$$

Hence

$$\gamma(u) \sim u\theta_u^2 \log(u) \text{ as } u \rightarrow \infty,$$

and substitution of θ_u from Theorem 5.3.2 gives the required result. \square

Attention now turns to estimation of the parameters for the DFARIMA model. The following proposition shows the work done by Fox and Taqqu [1986] and Dahlhaus [1989] still holds.

Proposition 5.3.1. *Let y_t be a DFARIMA process defined by equation 5.95 with spectral density $f_y(\lambda)$ defined by equation 5.96. Then the conditions of Fox and Taqqu [1986] and Dahlhaus [1989] for Theorem 2.3.1 hold for this process.*

Proof. Let

$$f_y(\lambda, \theta, c) = b(\lambda, c)f_x(\lambda, \theta),$$

where $f_x(\lambda, \theta)$ is assumed to satisfy the conditions of Fox and Taqqu [1986] and Dahlhaus [1989]. Then, $f_y(\lambda, \theta, c)$ will also satisfy the conditions of Fox and Taqqu [1986] and Dahlhaus [1989] if the following conditions on $b(\lambda, c)$ hold for $\eta > 0$ and some constants C, C_1 and C_2 which can be chosen independently of c but not of η .

B1) $b(\lambda, c)$ is continuous for all (λ, c) , $\lambda \neq 0$

$$C_1(\eta)\lambda^\eta < b(\lambda, c) < C_2(\eta)\lambda^{-\eta}$$

B2) $\partial/\partial c b(\lambda, c)$, $\partial^2/\partial c^2 b(\lambda, c)$ and $\partial^3/\partial c^3 b(\lambda, c)$ are continuous at all (λ, c) , $\lambda \neq 0$,

$$\left| \frac{\partial b(\lambda, c)}{\partial c} \right| \leq C(\eta)|\lambda|^{-\eta}$$

$$\left| \frac{\partial^2 b(\lambda, c)}{\partial c^2} \right| \leq C(\eta)|\lambda|^{-\eta}$$

and

$$\left| \frac{\partial^3 b(\lambda, c)}{\partial c^3} \right| \leq C(\eta)|\lambda|^{-\eta}$$

B3) $\partial/\partial \lambda b(\lambda, c)$ and $\partial^2/\partial \lambda^2 b(\lambda, c)$ are continuous at all (λ, c) , $\lambda \neq 0$,

$$\left| \frac{\partial b(\lambda, c)}{\partial \lambda} \right| \leq C(\eta)|\lambda|^{-1-\eta},$$

$$\left| \frac{\partial^2 b(\lambda, c)}{\partial \lambda^2} \right| \leq C(\eta)|\lambda|^{-2-\eta},$$

$$\left| \frac{\partial^2 b(\lambda, c)}{\partial \lambda \partial c} \right| \leq C(\eta)|\lambda|^{-1-\eta}$$

and

$$\left| \frac{\partial^3 b(\lambda, c)}{\partial \lambda^2 \partial c} \right| \leq C(\eta)|\lambda|^{-2-\eta}$$

B4) $\partial/\partial \lambda b(\lambda, c)^{-1}$ and $\partial^2/\partial \lambda^2 b(\lambda, c)^{-1}$ are continuous at all (λ, c) , $\lambda \neq 0$,

$$\left| \frac{\partial b(\lambda, c)^{-1}}{\partial \lambda} \right| \leq C(\eta)|\lambda|^{-1-\eta}$$

and

$$\left| \frac{\partial^2 b(\lambda, c)^{-1}}{\partial \lambda^2} \right| \leq C(\eta) |\lambda|^{-2-\eta}$$

B5) For all $\lambda \in (0, \pi)$ and $c \in (-\infty, \infty)$,

$$0 < a < b(\lambda, c) < h < \infty,$$

for some constants a and h .

Take $f_y(\lambda, \theta, c)$ as the spectral density of the new DFARIMA model, $f_x(\lambda, \theta)$ as the spectral density of a standard FARIMA model, which has previously be shown to satisfy these conditions, see Fox and Taqqu [1986] and Dahlhaus [1989], and $b(\lambda, c) = |1 - \log(1 - e^{-i\lambda})|^{-2c}$. Note, $b(\lambda, c)$ is continuous for $\lambda \in (0, \pi)$ and all its partial derivatives are bounded above and below away from $\lambda = 0$, thus B5 is satisfied and it need only be shown that $b(\lambda, c)$ satisfies assumptions B1 - B4 as $\lambda \rightarrow 0$. Also, $b(\lambda, c) \sim \log(1/\lambda)^{-2c}$ as $\lambda \rightarrow 0$. Since this is a slowly varying function, B1 is satisfied.

Now

$$\left| \frac{\partial^k b(\lambda, c)}{\partial c^k} \right| \sim \left| 2^k \log(\log(1/\lambda))^k \log(1/\lambda)^{-2c} \right| \leq C(\eta) |\lambda|^{-\eta} \text{ as } \lambda \rightarrow 0$$

thus B2 is satisfied.

Next,

$$\left| \frac{\partial b(\lambda, c)}{\partial \lambda} \right| \sim \left| \frac{2c \log(1/\lambda)^{-2c-1}}{\lambda} \right| \leq C(\eta) |\lambda|^{-1-\eta} \text{ as } \lambda \rightarrow 0,$$

$$\left| \frac{\partial^2 b(\lambda, c)}{\partial \lambda^2} \right| \sim \left| \frac{2c(2c+1) \log(1/\lambda)^{-2c-2} \lambda + 2c \log(1/\lambda)^{-2c-1}}{\lambda^2} \right| \leq C(\eta) |\lambda|^{-2-\eta} \text{ as } \lambda \rightarrow 0,$$

$$\left| \frac{\partial^2 b(\lambda, c)}{\partial \lambda \partial c} \right| \sim \left| \frac{2 \log(1/\lambda)^{-2c-1} - 4c \log(\log(1/\lambda)) \log(1/\lambda)^{-2c-1}}{\lambda} \right| \leq C(\eta) |\lambda|^{-1-\eta} \text{ as } \lambda \rightarrow 0,$$

and

$$\left| \frac{\partial^3 b(\lambda, c)}{\partial \lambda^2 \partial c} \right| \sim \left| \frac{(8c+2) \log(1/\lambda)^{-2c-2} - 2(4c^2+2c) \log(\log(1/\lambda)) \log(1/\lambda)^{-2c-2}}{\lambda} \right|$$

$$+ \left| \frac{2 \log(1/\lambda)^{-2c-1} - 4c \log(\log(1/\lambda)) \log(1/\lambda)^{-2c-1}}{\lambda^2} \right| \leq C(\eta) |\lambda|^{-2-\eta} \text{ as } \lambda \rightarrow 0,$$

thus B3 is satisfied. Finally, substitution of $b(\lambda, -c)$ into these equations shows B4 is satisfied. This completes the proof. \square

Proposition 5.3.1 shows the asymptotic theory of Fox and Taquq [1986], Dahlhaus [1989] and Giraitis and Surgailis [1990] still holds. The work of Fox and Taquq [1986] and Dahlhaus [1989] gives the asymptotic distributions of the exact maximum likelihood and Whittle maximum likelihood methods as stated in Theorem 2.3.1 under the assumption that y_t is Gaussian. For the Whittle likelihood estimates, the work of Giraitis and Surgailis [1990] relaxes the assumption of Gaussianity.

Note, for a DFARIMA(p,d,c,q) process, using the form of the spectral density given in equation 5.96 gives

$$\begin{aligned} \left(\frac{\partial \log(f_{\theta}(\lambda))}{\partial c} \right) = & -\log \left(\log \left(\frac{1}{2 \sin(\lambda/2)} \right)^2 + \log \left(\frac{e}{4 \sin(\lambda/2)^2} \right) \right) \\ & + \left(\tan^{-1} \left(\frac{\cos(\lambda/2)}{\sin(\lambda/2)} \right) \right)^2 \end{aligned} \quad (5.101)$$

Integration of this function with respect to λ is not straightforward and thus explicit forms for Γ given in Theorem 2.3.1 can be difficult to find. This problem can be overcome by use of Riemann sums, that is, by taking

$$\Gamma_{ij}(\theta) \approx \frac{1}{2\tilde{n}} \sum_{j=1}^{\tilde{n}} \left(\frac{\partial \log(f_{\theta}(\lambda_j))}{\partial \theta_i} \right) \left(\frac{\partial \log(f_{\theta}(\lambda_j))}{\partial \theta_j} \right),$$

where $\lambda_j = 2\pi j/\tilde{n}$ and \tilde{n} can be chosen as large as computationally convenient to improve accuracy. Now, for estimates from a DFARIMA(0,d,c,0) process, noting

$$\left(\frac{\partial \log(f_{\theta}(\lambda))}{\partial d} \right) = -2 \log \left(2 \sin \left(\frac{\lambda}{2} \right) \right) \quad (5.102)$$

use of the Riemann sums approximations, with $\tilde{n} = 10^7$, gives

$$\Gamma(d_0, c_0) \approx \begin{pmatrix} 1.64491 & -1.11786 \\ -1.11786 & 1.04235 \end{pmatrix}. \quad (5.103)$$

Note, an exact value of Γ_{11} is given by $\pi^2/6 = 1.64493$, thus the approximation of this value is correct to four decimal places, see for example Palma [2007] page 105. Use of Theorem 2.3.1 now gives the asymptotic distribution for estimates \hat{d} , \hat{c}

of d_0, c_0 from a DFARIMA(0, $d_0, c_0, 0$) process, found by maximising either the exact likelihood or the Whittle likelihood, as $n \rightarrow \infty$, as

$$\sqrt{n} \begin{pmatrix} \hat{d} - d_0 \\ \hat{c} - c_0 \end{pmatrix} \rightarrow_d N \left(0, \begin{pmatrix} 2.242 & 2.404 \\ 2.404 & 3.538 \end{pmatrix} \right). \quad (5.104)$$

Note, the asymptotic correlation between \hat{d} and \hat{c} is approximately equal to 0.85. Although some correlation is therefore present, the covariance matrix of the estimates is non-singular. Hence, the parametric estimates do not suffer from the same asymptotic colinearity as the non-parametric methods.

Note also that the asymptotic variances do not depend on the true values d_0 and c_0 . Since neither equation 5.101 or equation 5.102 depend on d_0 or c_0 , this will be true for all values of $p, q \in (0, 1, 2, \dots)$, however, as with the FARIMA model, dependence on $\theta_1, \dots, \theta_p$ and ϕ_1, \dots, ϕ_q will occur, see for example Palma [2007] page 106.

For two further examples, the estimate of d_0 for the DFARIMA(0, $d_0, 0, 0$) process, which is equivalent to a FARIMA(0, $d_0, 0$) process, has asymptotic distribution

$$\sqrt{n}(\hat{d} - d_0) \rightarrow_d N(0, 0.608), \text{ as } n \rightarrow \infty \quad (5.105)$$

see for example Palma [2007] page 104 and use of the Riemann sums again gives the asymptotic distribution of the estimate of c_0 for a DFARIMA(0, 0, $c_0, 0$) process as

$$\sqrt{n}(\hat{c} - c_0) \rightarrow_d N(0, 0.959), \text{ as } n \rightarrow \infty. \quad (5.106)$$

Due to standard 2×2 matrix inversion equations, the asymptotic variances for \hat{d} and \hat{c} when estimated simultaneously, given in equation 5.104, are both larger than the corresponding asymptotic variances when estimated separately, given in equations 5.105 and 5.106, by a factor of

$$\frac{\Gamma_{11}\Gamma_{22}}{\Gamma_{11}\Gamma_{22} - \Gamma_{21}^2} \approx \frac{(1.64)(1.04)}{(1.64)(1.04) - (-1.12)^2} = 3.69$$

where Γ_{ij} are the elements of the 2×2 matrix Γ for the DFARIMA(0, $d_0, c_0, 0$) model and the approximations come from the Riemann sums given in equation 5.103.

5.4 Simulation Study

In this section a simulation study is carried out to investigate the finite sample behaviour of the new methods of estimating d and c discussed in sections 5.1-5.3. The results are presented in two parts. Section 5.4.1 presents the results for the new non-parametric estimates found using the DGPH method discussed in section 5.1

and the dual local Whittle method presented in section 5.2. These are compared to the theoretical distributions given in Theorems 5.1.4 and 5.2.3 respectively. Results for the standard GPII and the standard local Whittle methods are also presented and the biases when $c \neq 0$ are compared to the theoretical asymptotic biases presented in Theorems 5.1.3 and 5.2.1. The results show a tendency towards the asymptotic results as the series length increases, however these non-parametric methods are shown to be unsuitable for shorter time series.

In section 5.4.2 results are presented from the parametric approach of fitting a DFARIMA(p, d, c, q) using the Whittle Likelihood method of estimation. Assuming p and q to be known, the finite sample distributions are compared to the theoretical asymptotic distributions given by Theorem 2.3.1. Assuming p and q to be unknown, AIC and BIC are used to find estimates of p and q from a range of possible values and the results are recorded. In particular, when the series is generated from a process with $c \neq 0$, both AIC and BIC tend to favour the fit of a DFARIMA(p, d, c, q) model to a standard FARIMA(p, d, q) model.

5.4.1 Empirical Results of the Non-Parametric Methods

In this section, the finite sampling properties of estimates of d and c from the standard GPII, standard local Whittle, dual parameter GPII and dual parameter local Whittle methods are investigated and compared in a simulation study. The purpose of this study is to discover if the asymptotic results for these methods given in sections 5.1 and 5.2 hold for time series of finite length. As such, the time series under study are simulated from what can be considered the ideal case, that is, from a linear Gaussian process with spectral density of the form

$$f(\lambda) = G_0 |\lambda|^{-2d_0} \left| \log \left(\frac{1}{\lambda} \right) \right|^{-2c_0} \quad \text{for } \lambda \in (-\pi, \pi). \quad (5.107)$$

These are simulated using the method proposed by Davis and Harte [1987], which is described in the following Theorem.

Theorem 5.4.1. *For $j \in (0, \dots, n/2)$, take z_j to be a sequence of independent complex normal random variables with independent real and imaginary parts. Let z_0 and $z_{n/2}$ be real with variances, $\sigma^2 = 2$, whilst the real and imaginary parts of z_j for $j \in (1, \dots, n/2 - 1)$ have variance $\sigma^2 = 1$. Finally, for $j \in (-n/2, \dots, -1)$, let $z_j = \bar{z}_{-j}$, where the bar above the z_{-j} represents the complex conjugate. Then, the series y_t , defined by*

$$y_t = \frac{1}{2n^{1/2}} \sum_{j=1-n/2}^{n/2} z_j f(\lambda_j)^{1/2} e^{i\lambda_j t/n}$$

for $t \in (1, \dots, n)$, has spectral density $f(\lambda)$.

Three series lengths are considered, $n_1 = 10^3$, $n_2 = 10^4$ and $n_3 = 10^5$. For each series length, time series are generated from the spectrum given in equation 5.107 with $d_0 \in (-0.5, -0.25, 0, 0.25, 0.5)$ and $c_0 \in (1, 0.5, 0, -0.5, -1)$. Finally, each of the (n_i, d_k, c_p) combinations was replicated 1000 times. It is noted that the asymptotic results given in sections 5.1 and 5.2 are not designed to allow for the cases $d_0 = 0.5$ when $c \leq 0$ or $d_0 = -0.5$ when $c \geq 0$, but they were retained in the simulation study to more fully examine the behaviour of the estimates at the boundary conditions.

For each simulated series the standard GPII and standard local Whittle methods were used to find estimates of d_0 and the new dual parameter GPII and dual parameter local Whittle methods were used to find estimates of d_0 and c_0 for three different values of m , as described in sections 5.1 and 5.2. For simulated series of length n_1 , the three values of m were $m_{11} = 10$, $m_{12} = 30$ and $m_{13} = 50$, for simulated series of length n_2 , the three values of m were $m_{21} = 100$, $m_{22} = 150$ and $m_{23} = 200$ and for simulated series of length n_3 , the three values of m were $m_{31} = 500$, $m_{32} = 1000$ and $m_{33} = 1500$.

Since there is a zero or singularity in the spectral density at $\lambda = 1$, these values of m were chosen to avoid the effect of this point by taking $\lambda_m < 1/3$. Also, since it is assumed $m/n \rightarrow 0$ as $n \rightarrow \infty$, the values of m were chosen to reduce the proportion of frequencies used for longer time series.

For the GPII-type methods a lower point of truncation for frequencies used, λ_l , should also be chosen, see section 5.1. However, theoretical work done by Hurvich, Deo and Brodsky [1998] shows for the standard GPII method when $c = 0$, l can be taken as equal to 1 with negligible asymptotic effect, whilst simulation studies by Hurvich and Beltra [1994] show for the standard GPII method, when $c = 0$, taking $l > 1$ leads to finite sample estimates with increased variance and only a very small reduction in bias. Using various values of l , it was also found here that no significant advantages for either the standard GPII method or the dual parameter GPII method were gained by taking $l > 1$ when $c \neq 0$ and it is believed that the theoretical work done by Hurvich, Deo and Brodsky [1998] can be extended to include the case when $c \neq 0$, though this is left for future work. Due to these arguments, and for comparison to the local Whittle-type estimates, l is taken as equal to 1 throughout for the results presented in this section.

The results for the standard methods of estimating d_0 are presented first. A discussion of the finite sampling properties for these standard methods can be found in section 4.1.

The standard deviations for \hat{d} estimated using the standard local Whittle and standard GPII methods with a series length n_3 and bandwidth m_{32} are shown in table 5.1. In general, the standard deviations seem to show no dependence on the true value of d_0 or c_0 and agree with the asymptotic results of Robinson [1995a] and Robinson [1995b]. Since no new theoretical work has been presented in sections 5.1 and 5.2 regarding the standard deviations of the estimates of d_0 from the standard GPII and standard local Whittle methods, the standard deviations for these methods

SD for standard LW estimates of d						SD for standard GPH estimates of d				
Series Length $n=100000$, Bandwidth $m=1000$										
Asymptotic SD= 0.016						Asymptotic SD= 0.020				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.104	0.034	0.017	0.017	0.017	0.026	0.021	0.022	0.021	0.020
0.5	0.031	0.017	0.017	0.016	0.017	0.022	0.022	0.021	0.021	0.021
0	0.017	0.017	0.018	0.017	0.017	0.022	0.022	0.022	0.021	0.021
-0.5	0.017	0.017	0.018	0.017	0.017	0.022	0.021	0.021	0.021	0.021
-1	0.017	0.017	0.017	0.017	0.017	0.022	0.021	0.021	0.021	0.021

Tab. 5.1: Sample standard deviations for \hat{d} estimated using the standard local Whittle and standard GPH methods with a series length of $n = 100000$ and a bandwidth of $m = 1000$.

for the remainder of (n, m) combinations have been omitted to save space. Suffice to say, the standard deviations agree by in large with the previous simulation studies mentioned in that the shorter series lengths of n_1 have slightly larger standard deviations than the asymptotic theory suggests, though the asymptotic theory seems to hold very well for lengths n_2 and n_3 .

Interesting, however, is the behaviour of the standard deviations for $d_0 = -0.5$ and $c_0 > 0$. Table 5.1 shows the sample standard deviations for the estimates of d_0 from the standard local Whittle method for $d_0 = -0.5$ and $c_0 > 0$ are significantly larger than the other values of d_0 and c_0 and, to a lesser extent, the same is true for the standard GPH method. This is visible for every (n, m) combination considered in this simulation study.

Theorem 3.2 of Phillips and Shimotsu [2006] shows for $d_0 < -0.5$, as $n \rightarrow \infty$ the local Whittle estimate will converge either to the true value of d_0 or to 0 depending on the choice of m . The simulation study presented in Phillips and Shimotsu [2006] shows in practice this tends to lead to larger standard deviations and large positive biases when $d_0 < -0.5$. Hurvich and Ray [1995] give theoretical results to show estimates of d_0 found via the standard GPH method also tend to zero as the series length increases for $c_0 = 0$ and $d_0 < -0.5$. The simulation study presented in Hurvich and Ray [1995] shows in practice this also seems to lead to larger standard deviations and large positive biases for estimates of d_0 from the standard GPH method.

Since a positive value of c_0 causes a decrease in the memory of the process, the case when $d_0 = -0.5$ and $c_0 > 0$ is on the upper boundary of the range of d_0 covered by Theorem 3.2 of Phillips and Shimotsu [2006] and Hurvich and Ray [1995] and the results presented in table 5.1 suggest these theoretical results may still hold in some form at this boundary condition.

The biases for the standard local Whittle method are presented in table 5.2, although in the interest of space only the results for the second value of m for each

Bias for standard Local Whittle estimates of d						theoretical Bias				
Series Length $n=1000$, Bandwidth $m=30$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.280	-0.045	-0.270	-0.370	-0.400	-0.45	-0.45	-0.45	-0.45	-0.45
0.5	-0.019	-0.150	-0.200	-0.220	-0.240	-0.22	-0.22	-0.22	-0.22	-0.22
0	0.036	-0.019	-0.045	-0.053	-0.068	0.00	0.00	0.00	0.00	0.00
-0.5	0.170	0.140	0.130	0.120	0.110	0.22	0.22	0.22	0.22	0.22
-1	0.340	0.320	0.300	0.290	0.261	0.45	0.45	0.45	0.45	0.45
Series Length $n=10000$, Bandwidth $m=150$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.130	-0.142	-0.264	-0.270	-0.280	-0.33	-0.33	-0.33	-0.33	-0.33
0.5	-0.058	-0.131	-0.141	-0.140	-0.150	-0.16	-0.16	-0.16	-0.16	-0.16
0	0.013	-0.003	-0.008	-0.017	-0.018	0.00	0.00	0.00	0.00	0.00
-0.5	0.140	0.133	0.120	0.110	0.110	0.16	0.16	0.16	0.16	0.16
-1	0.260	0.262	0.252	0.250	0.240	0.33	0.33	0.33	0.33	0.33
Series Length $n=100000$, Bandwidth $m=1000$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.011	-0.190	-0.230	-0.230	-0.234	-0.26	-0.26	-0.26	-0.26	-0.26
0.5	-0.084	-0.110	-0.120	-0.123	-0.121	-0.13	-0.13	-0.13	-0.13	-0.13
0	0.004	-0.000	-0.002	-0.003	-0.003	0.00	0.00	0.00	0.00	0.00
-0.5	0.120	0.110	0.110	0.114	0.114	0.13	0.13	0.13	0.13	0.13
-1	0.230	0.232	0.232	0.231	0.225	0.26	0.26	0.26	0.26	0.26

Tab. 5.2: Sample and theoretical biases for \hat{d} estimated using the standard local Whittle method.

series length are presented. For a fixed series length, the bias increases with m when $c_0 \neq 0$, as expected from Theorem 5.2.1. The theoretical biases are obtained from the upper and lower bounds presented in Theorem 5.2.1. Since Theorem 5.2.1 gives bounds for the expected bias rather than the biases in an explicit form, it is expected that the modulus of the sample biases be smaller than those of the theoretical biases. Table 5.2 shows this is generally the case and the same holds for those results not shown.

When $c_0 = 0$ the results seem to agree with those presented by the previous simulation studies already mentioned, that is the biases are small in magnitude, with a tendency towards being negative and decrease as the series length n increases. When $c_0 \neq 0$, however, table 5.2 shows the biases seem to be close to the upper and lower bounds presented in Theorem 5.2.1. For series lengths of $n \geq 10^4$ in particular, the biases should generally be considered unacceptable. Comparison with the standard deviations given in table 5.1 shows the biases when $c_0 \neq 0$, although decreasing with n , are generally larger than the standard deviations for $n = 10^5$, $m = 10^3$. This is also true for the other (n, m) combinations included in this study, implying tests on the parameters may lead to incorrect conclusions.

The case when $d_0 = -0.5$ and $c_0 > 0$ again stands out. Far from the strongly negative bias which would be expected if Theorem 5.2.1 were to hold, the biases tend to be positive, suggesting again that some form of Theorem 3.2 of Phillips and Shimotsu [2006] may still hold for this boundary condition and that Theorem 5.2.1 does not hold under these conditions. Interestingly, however, table 5.2 suggests Theorem 5.2.1 may still hold for $d_0 = 0.5$ with $c_0 \leq 0$, which is no longer stationary.

The biases, for the second value of m for each series length, for the standard GPII method are presented in table 5.3. The remaining results are omitted in the interest of space, though it is noted, as with the standard local Whittle estimates, the bias increases with m when $c_0 \neq 0$. This is as expected from Theorem 5.1.3.

As mentioned in remark 5.1.3, the rate of convergence to the asymptotic results given in Theorem 5.1.3 is expected to be slow and in practise it is preferable to calculate the actual theoretical biases rather than the asymptotic limits. The theoretical biases given in Table 5.3 are thus the actual theoretical biases, not the asymptotic limits given by Theorem 5.1.3.

Generally, the biases for the estimates from the standard GPII method are similar in magnitude to those for the standard local Whittle estimates. As for the estimates for the standard local Whittle method, when $c_0 = 0$ the biases are small in magnitude and decrease as the series length n increases. This is in agreement with the results presented by the previous simulation studies already mentioned. When $c_0 \neq 0$, table 5.3 shows the biases seem close to the theoretical ones, the asymptotic limits of which are given in Theorem 5.2.1. The magnitude of these biases should generally be considered unacceptable for series lengths of $n \geq 10^4$. As with the standard local Whittle estimates, comparison with the standard deviations given in table 5.1 shows the biases when $c_0 \neq 0$ are generally larger than the standard deviations for $n = 10^5$,

Bias for standard GPH estimates of d						theoretical Bias				
Series Length $n=1000$, Bandwidth $m=30$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.163	-0.094	-0.276	-0.358	-0.388	-0.35	-0.35	-0.35	-0.35	-0.35
0.5	-0.032	-0.138	-0.187	-0.215	-0.227	-0.17	-0.17	-0.17	-0.17	-0.17
0	0.051	-0.009	-0.034	-0.052	-0.067	0.00	0.00	0.00	0.00	0.00
-0.5	0.184	0.150	0.135	0.125	0.110	0.17	0.17	0.17	0.17	0.17
-1	0.346	0.319	0.304	0.281	0.242	0.35	0.35	0.35	0.35	0.35
Series Length $n=10000$, Bandwidth $m=150$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	-0.080	-0.198	-0.257	-0.264	-0.276	-0.26	-0.26	-0.26	-0.26	-0.26
0.5	-0.085	-0.125	-0.135	-0.140	-0.144	-0.13	-0.13	-0.13	-0.13	-0.13
0	0.016	-0.001	-0.004	-0.017	-0.018	0.00	0.00	0.00	0.00	0.00
-0.5	0.138	0.129	0.119	0.113	0.111	0.13	0.13	0.13	0.13	0.13
-1	0.263	0.256	0.249	0.246	0.230	0.26	0.26	0.26	0.26	0.26
Series Length $n=100000$, Bandwidth $m=1000$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	-0.183	-0.218	-0.229	-0.231	-0.234	-0.23	-0.23	-0.23	-0.23	-0.23
0.5	-0.103	-0.114	-0.117	-0.117	-0.118	-0.11	-0.11	-0.11	-0.11	-0.11
0	0.004	0.001	-0.002	-0.004	-0.004	0.00	0.00	0.00	0.00	0.00
-0.5	0.117	0.114	0.112	0.111	0.110	0.11	0.11	0.11	0.11	0.11
-1	0.230	0.228	0.227	0.224	0.220	0.23	0.23	0.23	0.23	0.23

Tab. 5.3: Sample and theoretical biases for \hat{d} estimated using the standard GPH method.

$m = 10^3$. Again, this is also true for the other (n, m) combinations included in this study, once more implying tests on the parameters may lead to incorrect conclusions.

For the case when $d_0 = -0.5$ and $c_0 > 0$, the biases of the estimates of d_0 using the standard GPH are not as negative as would be expected if the theoretical results which give the asymptotic theory of Theorem 5.1.3 were to hold. This once again suggests some form of the theoretical results of Hurvich and Ray [1995] may still hold for this boundary condition. However, as with the standard deviations presented in table 5.1, the extent to which the results of Hurvich and Ray [1995] seem to hold for the standard GPH method seem less than the extent to which Theorem 3.2 of Phillips and Shimotsu [2006] seems to hold for the standard local Whittle method.

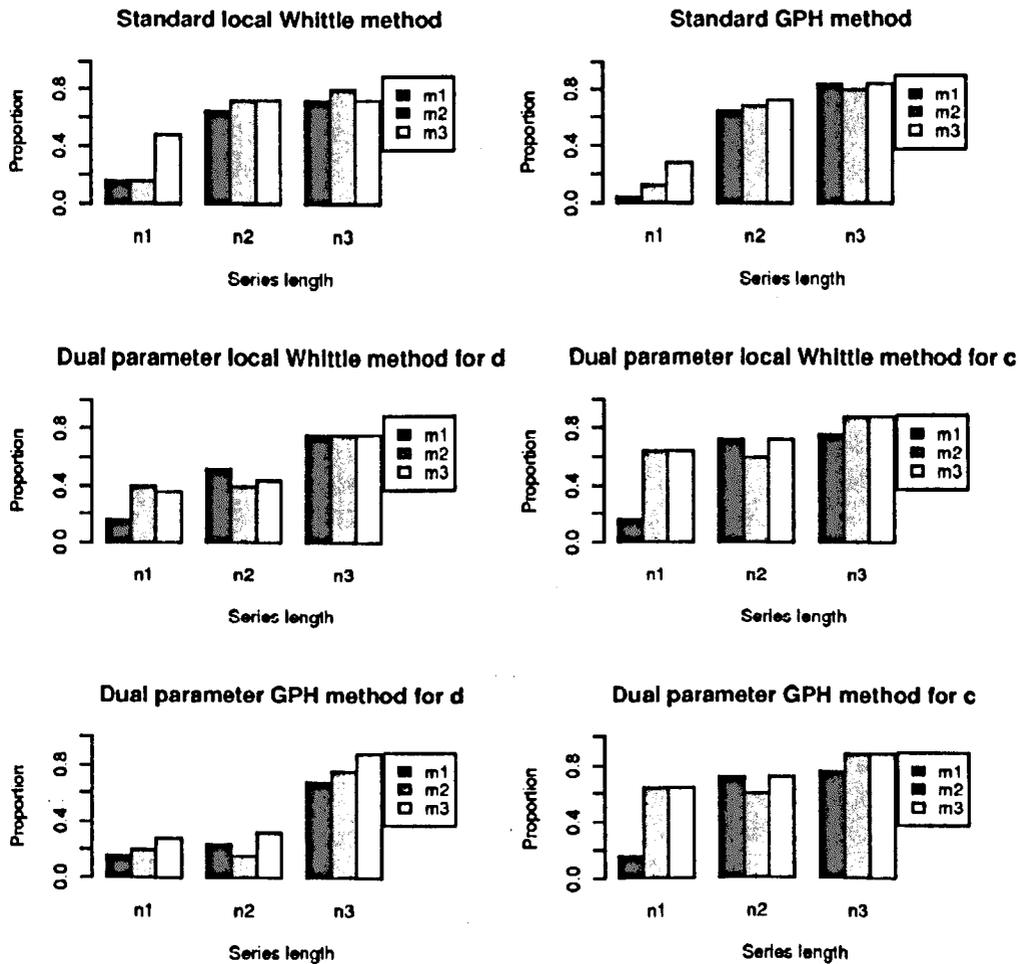


Fig. 5.2: The proportion of sampling distributions which were not significantly different from the normal distribution at the 5% significance level for each series length and bandwidth.

Bias for standard GPH estimates of d						theoretical Bias				
Series Length $n=1000$, Bandwidth $m=30$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.163	-0.094	-0.276	-0.358	-0.388	-0.35	-0.35	-0.35	-0.35	-0.35
0.5	-0.032	-0.138	-0.187	-0.215	-0.227	-0.17	-0.17	-0.17	-0.17	-0.17
0	0.051	-0.009	-0.034	-0.052	-0.067	0.00	0.00	0.00	0.00	0.00
-0.5	0.184	0.150	0.135	0.125	0.110	0.17	0.17	0.17	0.17	0.17
-1	0.346	0.319	0.304	0.281	0.242	0.35	0.35	0.35	0.35	0.35
Series Length $n=10000$, Bandwidth $m=150$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	-0.080	-0.198	-0.257	-0.264	-0.276	-0.26	-0.26	-0.26	-0.26	-0.26
0.5	-0.085	-0.125	-0.135	-0.140	-0.144	-0.13	-0.13	-0.13	-0.13	-0.13
0	0.016	-0.001	-0.004	-0.017	-0.018	0.00	0.00	0.00	0.00	0.00
-0.5	0.138	0.129	0.119	0.113	0.111	0.13	0.13	0.13	0.13	0.13
-1	0.263	0.256	0.249	0.246	0.230	0.26	0.26	0.26	0.26	0.26
Series Length $n=100000$, Bandwidth $m=1000$										
	d					d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	-0.183	-0.218	-0.229	-0.231	-0.234	-0.23	-0.23	-0.23	-0.23	-0.23
0.5	-0.103	-0.114	-0.117	-0.117	-0.118	-0.11	-0.11	-0.11	-0.11	-0.11
0	0.004	0.001	-0.002	-0.004	-0.004	0.00	0.00	0.00	0.00	0.00
-0.5	0.117	0.114	0.112	0.111	0.110	0.11	0.11	0.11	0.11	0.11
-1	0.230	0.228	0.227	0.224	0.220	0.23	0.23	0.23	0.23	0.23

Tab. 5.3: Sample and theoretical biases for d estimated using the standard GPH method.

$m = 10^3$. Again, this is also true for the other (n, m) combinations included in this study, once more implying tests on the parameters may lead to incorrect conclusions.

For the case when $d_0 = -0.5$ and $c_0 > 0$, the biases of the estimates of d_0 using the standard GPH are not as negative as would be expected if the theoretical results which give the asymptotic theory of Theorem 5.1.3 were to hold. This once again suggests some form of the theoretical results of Hurvich and Ray [1995] may still hold for this boundary condition. However, as with the standard deviations presented in table 5.1, the extent to which the results of Hurvich and Ray [1995] seem to hold for the standard GPH method seem less than the extent to which Theorem 3.2 of Phillips and Shimotsu [2006] seems to hold for the standard local Whittle method.

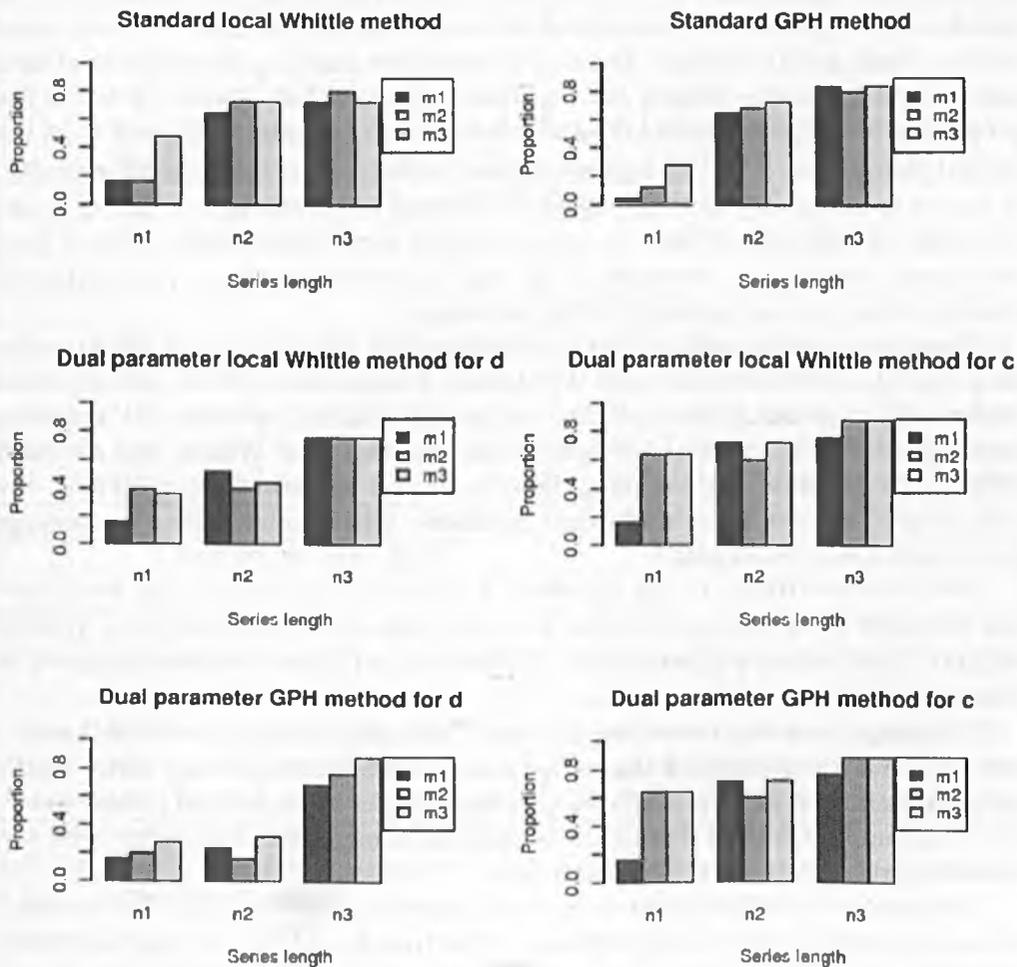


Fig. 5.2: The proportion of sampling distributions which were not significantly different from the normal distribution at the 5% significance level for each series length and bandwidth.

Since the theoretical asymptotic distributions are normal for the estimates of d_0 and c_0 from the standard local Whittle, standard GPII, dual parameter local Whittle and dual parameter GPII methods a test for normality was carried out on each of the sampling distributions for each (d_0, c_0, n_i, m_{ij}) combination. The null hypothesis for the tests was normality of the sampling distributions vs. the alternative hypothesis of non-normality. The tests used Shapiro-Wilk's W statistic, see Shapiro and Wilk [1965], which was shown to be generally superior in detecting non-normality than several other statistics in a comparative study by Shapiro, Wilk and Chen [1968].

Generally, the p-value of the normality tests did not depend on the values of d_0 and c_0 . The exception to this was when $d_0 = -0.5$ with $c_0 = 1$, as for this case the sampling distributions of the estimates of d_0 and c_0 for all four methods proved significantly different to the normal distribution for all series lengths and bandwidths chosen. Disregarding this case, there are therefore 24 sampling distributions of each estimate for each series length and bandwidth combination. Figure 5.2 shows the proportion of sampling distributions which were not significantly different from the normal distribution at the 5% significance level for each series length and bandwidth. It can be seen that as the series length n increases the proportion of sampling distributions for estimates of both d_0 and c_0 which were not significantly different from the normal distribution increases for all four estimation methods, reinforcing the theoretical asymptotic normality of the estimates.

There is also evidence to suggest that the sampling distributions for the estimates of c_0 from the dual parameter local Whittle and dual parameter GPII methods seem to converge to normality at a similar, and possibly faster, rate than the sampling distributions for the estimates of d_0 from the standard local Whittle and standard GPII methods, whilst the sampling distributions for the estimates of d_0 from the dual parameter local Whittle and dual parameter GPII methods seem to converge to normality at a slower rate.

Attention now turns to the standard deviations and biases of the simultaneous estimates of d_0 and c_0 using the new dual parameter GPII and local Whittle methods. The theoretical asymptotic distributions for these estimates are given in Theorems 5.1.4 and 5.2.3.

The sample standard deviations for \hat{d} and \hat{c} estimated using the new dual parameter local Whittle method for the second choice of bandwidth for each series length are given in table 5.4. The results for the other choices of m , omitted to save space, show the sample standard deviations decrease as m increases. This agrees with the theoretical results given by Theorem 5.2.3.

As expected by the asymptotic distribution given in Theorem 5.2.3, the standard deviations generally show no dependence of the true d_0 and c_0 . The exception once again being for the case when $d_0 = -0.5$ and $c_0 > 0$. Whilst the observation of the increased standard deviations for the standard local Whittle estimates at these points suggested some form of Theorem 3.2 of Phillips and Shimotsu [2006] may still hold, the observation of increased standard deviations for the dual parameter estimates at these points suggests Theorem 3.2 of Phillips and Shimotsu [2006] may

SD for DLW estimates of d						SD for DLW estimates of c				
Series Length n=1000, Bandwidth m =30										
Asymptotic SD= 0.315						Asymptotic SD= 2.178				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	1.058	0.831	0.771	0.770	0.767	2.718	2.283	2.175	2.152	2.148
0.5	0.933	0.740	0.768	0.775	0.787	2.492	2.097	2.178	2.172	2.208
0	0.802	0.820	0.800	0.772	0.765	2.233	2.329	2.243	2.174	2.150
-0.5	0.791	0.818	0.817	0.796	0.810	2.238	2.300	2.277	2.228	2.257
-1	0.798	0.805	0.785	0.804	0.798	2.251	2.233	2.190	2.213	2.208
Series Length n=10000, Bandwidth m =150										
Asymptotic SD= 0.188						Asymptotic SD= 1.732				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.569	0.370	0.308	0.300	0.294	1.929	1.336	1.164	1.136	1.099
0.5	0.429	0.294	0.302	0.310	0.311	1.535	1.105	1.138	1.171	1.180
0	0.310	0.303	0.297	0.313	0.305	1.171	1.141	1.129	1.173	1.143
-0.5	0.314	0.298	0.305	0.308	0.315	1.179	1.117	1.143	1.148	1.183
-1	0.316	0.309	0.312	0.322	0.326	1.185	1.150	1.162	1.210	1.202
Series Length n=100000, Bandwidth m =1000										
Asymptotic SD= 0.091						Asymptotic SD= 1.048				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.329	0.164	0.109	0.110	0.101	1.285	0.675	0.465	0.471	0.432
0.5	0.155	0.105	0.104	0.106	0.108	0.640	0.449	0.452	0.455	0.464
0	0.106	0.104	0.109	0.109	0.111	0.460	0.446	0.469	0.471	0.479
-0.5	0.106	0.107	0.111	0.108	0.107	0.455	0.460	0.480	0.465	0.459
-1	0.106	0.109	0.108	0.109	0.110	0.457	0.471	0.466	0.466	0.472

Tab. 5.4: Sample standard deviations for \hat{d} and \hat{c} estimated using the new dual parameter local Whittle method.

also be able to be generalised to give similar results for the dual parameter local Whittle method.

The magnitudes of the standard deviations for the dual parameter local Whittle estimates shown in table 5.4 are of some concern. Although decreasing as the series length n increases, as suggested by Theorem 5.2.3, even for series lengths of $n = 10000$ the sample standard deviations for the estimates of d_0 seem to be around 0.3 and since d_0 is often in practise assumed to be in a range such as $[0, 0.5]$, the size of the standard deviation implies no meaningful conclusions can be made. Also of concern, the sample standard deviations seem to converge very slowly to the asymptotic results. The sample standard deviations for the estimates of d_0 are all larger than the asymptotic theory suggests. Interestingly, the sample standard deviations for the estimates of c_0 are often smaller than the asymptotic theory suggests

SD for DGPH estimates of d						SD for DGPH estimates of c				
Series Length n=1000, Bandwidth m =30										
theoretical SD= 0.873						theoretical SD= 2.495				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	1.179	0.920	0.912	0.873	0.901	3.084	2.580	2.587	2.495	2.552
0.5	1.047	0.854	0.881	0.926	0.934	2.854	2.458	2.519	2.604	2.640
0	0.931	0.930	0.908	0.921	0.882	2.623	2.666	2.580	2.629	2.522
-0.5	0.910	0.917	0.938	0.914	0.899	2.599	2.610	2.659	2.584	2.544
-1	0.920	0.912	0.900	0.914	0.916	2.637	2.577	2.563	2.583	2.573
Series Length n=10000, Bandwidth m =150										
theoretical SD= 0.350						theoretical SD= 1.329				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.537	0.390	0.359	0.366	0.356	1.892	1.447	1.373	1.401	1.350
0.5	0.435	0.354	0.371	0.373	0.360	1.597	1.343	1.405	1.414	1.376
0	0.379	0.365	0.362	0.374	0.373	1.437	1.384	1.379	1.416	1.400
-0.5	0.383	0.367	0.371	0.368	0.379	1.445	1.395	1.406	1.388	1.426
-1	0.381	0.354	0.381	0.368	0.388	1.438	1.332	1.434	1.393	1.449
Series Length n=100000, Bandwidth m =1000										
theoretical SD= 0.127						theoretical SD= 0.550				
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.195	0.145	0.134	0.132	0.129	0.807	0.620	0.580	0.568	0.557
0.5	0.144	0.131	0.128	0.131	0.133	0.616	0.559	0.557	0.564	0.571
0	0.131	0.131	0.134	0.134	0.130	0.565	0.564	0.576	0.582	0.562
-0.5	0.129	0.131	0.138	0.134	0.130	0.559	0.563	0.601	0.581	0.557
-1	0.131	0.136	0.134	0.133	0.132	0.566	0.587	0.582	0.571	0.567

Tab. 5.5: Sample standard deviations for \hat{d} and \hat{c} estimated using the new DGPH method.

The sample standard deviations for \hat{d} and \hat{c} estimated using the new dual parameter GPII method for the second choice of bandwidth for each series length are given in table 5.5. As with the dual parameter local Whittle method results for the other choices of m show the sample standard deviations decrease as m increases and are omitted to save space. This agrees with the theoretical results given by Theorem 5.1.4.

The magnitudes of the standard deviations for the dual parameter GPII estimates shown in table 5.4 are also of some concern. As with the comparison between the standard GPII and standard local Whittle methods, the sample standard deviations for the dual parameter GPII method seem to be slightly larger whilst decreasing at a similar rate to those for the dual parameter local Whittle method. A comparison of Theorem 5.1.4 with Theorem 5.2.3 shows asymptotically the standard deviations for the dual parameter GPII method should be approximately 1.8 times the size of those for the dual parameter local Whittle method. A comparison of tables 5.4 and 5.5 shows the sample standard deviations tend to be approximately 1.2 times larger for the dual parameter GPII method compared with the dual parameter local Whittle method, regardless of d_0 , c_0 or series length.

The theoretical standard deviations given in table 5.5 seem to fit the sampling standard deviations very well, though it is noted these are not the asymptotic standard deviations but the theoretical standard deviations evaluated exactly, as suggested in remark 5.1.3. Also note once again that as expected by Theorem 5.1.4, the standard deviations generally show no dependence of the true d_0 and c_0 . The exception once more being for the case when $d_0 = -0.5$ and $c_0 > 0$, where the standard deviations are larger. This suggests the results of Hurvich and Ray [1995] may also be able to be generalised to give similar results for the dual parameter GPII method. However, as with the standard methods, the effect of these points on the standard deviations for the dual parameter GPII estimates seems smaller than that on the standard deviations for the dual parameter local Whittle estimates.

The biases for the estimates of d_0 and c_0 from the dual parameter local Whittle method are shown in table 5.6. Comparison with the biases for the estimates of d_0 from the standard local Whittle method, shown in table 5.2 show for series lengths of $n \geq 10^4$ there is generally a significant reduction in bias when $c_0 \neq 0$ for the dual parameter local Whittle estimates.

For the series length $n = 1000$, the biases of the estimates of d_0 are generally similar in modulus or the standard local Whittle estimates are superior. Since the standard deviations are also much smaller for the standard local Whittle estimates when $n = 1000$, the dual parameter local Whittle method seems unsuitable for this series length.

For a series length $n = 10,000$ it is difficult to suggest which method is superior. There is a trade off between the large biases of the standard local Whittle estimates when $c \neq 0$ and the large standard deviations of the dual parameter local Whittle estimates and thus both methods may easily lead to incorrect conclusions during hypothesis testing.

Bias for DLW estimates of d						Bias for DLW estimates of c				
Series Length $n=1000$, Bandwidth $m=30$										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	2.502	1.238	0.223	-0.250	-0.412	5.782	3.141	0.856	-0.363	-0.702
0.5	0.633	0.000	-0.288	-0.356	-0.481	1.239	0.390	-0.349	-0.401	-0.903
0	0.063	-0.172	-0.326	-0.477	-0.492	0.385	-0.124	-0.645	-0.849	-1.061
-0.5	-0.103	-0.271	-0.370	-0.492	-0.538	-0.043	-0.435	-0.667	-0.901	-1.152
-1	-0.144	-0.315	-0.401	-0.452	-0.660	-0.112	-0.522	-0.723	-0.821	-1.572
Series Length $n=10000$, Bandwidth $m=150$										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	1.558	0.519	-0.016	-0.114	-0.136	4.865	1.562	-0.092	-0.236	-0.411
0.5	0.378	0.014	-0.079	-0.116	-0.145	1.324	0.324	-0.136	-0.267	-0.427
0	0.086	-0.060	-0.078	-0.141	-0.171	0.564	-0.013	-0.256	-0.392	-0.521
-0.5	-0.002	-0.054	-0.103	-0.137	-0.164	-0.023	-0.134	-0.219	-0.387	-0.531
-1	-0.020	-0.084	-0.118	-0.146	-0.234	-0.124	-0.199	-0.392	-0.399	-0.772
Series Length $n=100000$, Bandwidth $m=1000$										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.848	0.142	-0.014	-0.029	-0.042	2.128	0.421	-0.019	-0.106	-0.145
0.5	0.154	0.002	-0.021	-0.028	-0.037	0.421	0.009	-0.059	-0.099	-0.147
0	0.036	-0.008	-0.024	-0.036	-0.039	0.159	-0.026	-0.115	-0.124	-0.152
-0.5	0.020	-0.008	-0.022	-0.035	-0.036	0.128	-0.043	-0.092	-0.121	-0.156
-1	0.003	-0.015	-0.027	-0.036	-0.060	0.027	-0.101	-0.128	-0.141	-0.224

Tab. 5.6: Sample biases for \hat{d} and \hat{c} estimated using the new dual parameter local Whittle method. theoretical biases are all zero.

For the series length $n = 10^5$ it is suggested that the dual parameter local Whittle method is preferred. When $c_0 \neq 0$, the standard local Whittle method leads to estimates with large biases and small variances, suggesting a high probability of incorrect conclusions. However, the biases of the dual parameter local Whittle estimates are much smaller when $c \neq 0$ and the standard deviations are of an acceptable size.

Bias for DGPH estimates of d						Bias for DGPH estimates of c				
Series Length n=1000, Bandwidth m =30										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	2.746	1.437	0.404	-0.130	-0.282	6.480	3.430	0.969	-0.342	-0.693
0.5	0.795	0.175	-0.135	-0.180	-0.343	1.890	0.406	-0.350	-0.399	-0.836
0	0.212	-0.011	-0.195	-0.333	-0.380	0.465	-0.004	-0.465	-0.814	-0.906
-0.5	0.065	-0.125	-0.237	-0.350	-0.419	0.156	-0.295	-0.574	-0.875	-1.030
-1	0.005	-0.196	-0.257	-0.325	-0.560	0.012	-0.492	-0.622	-0.755	-1.323
Series Length n=10000, Bandwidth m =150										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	1.390	0.495	0.039	-0.092	-0.118	4.650	1.670	0.140	-0.339	-0.391
0.5	0.356	0.058	-0.043	-0.083	-0.128	1.200	0.202	-0.146	-0.280	-0.438
0	0.127	-0.016	-0.053	-0.107	-0.144	0.427	-0.057	-0.190	-0.345	-0.485
-0.5	0.037	-0.015	-0.073	-0.101	-0.144	0.114	-0.057	-0.235	-0.322	-0.479
-1	0.025	-0.055	-0.096	-0.123	-0.216	0.088	-0.195	-0.328	-0.419	-0.714
Series Length n=100000, Bandwidth m =1000										
d						d				
c	-0.5	-0.25	0	0.25	0.5	-0.5	-0.25	0	0.25	0.5
1	0.456	0.100	-0.003	-0.020	-0.038	1.800	0.396	-0.008	-0.076	-0.144
0.5	0.098	0.008	-0.014	-0.025	-0.038	0.382	0.033	-0.049	-0.097	-0.149
0	0.040	0.005	-0.016	-0.032	-0.038	0.162	0.017	-0.062	-0.124	-0.151
-0.5	0.033	0.001	-0.018	-0.034	-0.037	0.133	0.003	-0.069	-0.135	-0.145
-1	0.010	-0.006	-0.023	-0.035	-0.067	0.037	-0.028	-0.096	-0.136	-0.257

Tab. 5.7: Sample biases for \hat{d} and \hat{c} estimated using the new DGPH method. theoretical biases are all zero.

The biases for the estimates of d_0 and c_0 from the dual parameter GPH method are shown in table 5.7. Comparison with the biases for the estimates of d_0 from the standard GPH method, shown in table 5.3 gives similar conclusions as the comparison between the standard local Whittle and the dual parameter local Whittle estimates.

For the series length $n = 1000$, due to unacceptably large standard deviations combined with no noticeable improvement of biases, the dual parameter GPH method seems unsuitable.

For a series length $n = 10,000$, there is a trade off between the large biases of the

standard GPH estimates when $c \neq 0$ and the large standard deviations of the dual parameter GPH estimates. As before, both methods may easily lead to incorrect conclusions during hypothesis testing.

For the series length $n = 10^5$ it is suggested that the dual parameter GPH method is preferred, due to the large reduction in biases of the dual parameter GPH estimates when $c \neq 0$ and the more acceptable sizes of the standard deviations.

Finally, comparison between the biases for the estimates of d_0 and c_0 from the dual parameter local Whittle method, shown in table 5.6, and the biases from the dual parameter GPH method, shown in table 5.7, show little difference between the methods. Since the standard deviations of the dual parameter local Whittle method are generally slightly smaller, this seems the preferred method of the two.

In conclusion, the results of the simulation study presented here reinforce the theoretical work done in sections 5.1 and 5.2 which suggest when $c \neq 0$ the standard non-parametric methods commonly used in the literature can lead to strongly biased results. Unfortunately, the new dual parameter methods designed to reduce this bias can have unacceptably large standard deviations for series lengths of $n \leq 10^4$. However, for series lengths of $n = 10^5$ the new methods appear to lead to superior estimates and the asymptotic results given in sections 5.1 and 5.2 suggest this advantage to increase further for longer time series. With increasingly high frequency sampling of time series such as internet traffic, series of the lengths required become increasingly common.

5.4.2 Empirical Results for Fitting DFARIMA Models

In this section results are presented for estimating d_0 and c_0 using the parametric approach of fitting a DFARIMA(p, d_0, c_0, q) model to the observed time series via maximising the Whittle likelihood approximation. Several previous simulation studies have been carried out using FARIMA(p, d_0, q) models, which are a special case of the new DFARIMA(p, d_0, c_0, q).

Taqqu, Teverovsky and Willinger [1995] give results on times series of length $n = 10,000$ generated from Gaussian FARIMA($0, d_0, 0$) processes and FGN with various values of d_0 . They show the Whittle likelihood estimates to out perform all of the non-parametric methods they consider, with typical sample biases of around 0.001 in modulus and sample standard deviations less than 0.01.

Taqqu and Teverovsky [1996] extend on these results including also FARIMA(p, d_0, q) processes with $p, q \in (0, 1)$ and in addition to finding the Whittle likelihood estimates of the parameters from the correct model, they study also the effect of under and over fitting. For a series length $n = 10,000$, they find once again that fitting the correct model leads to estimates of small sample bias and sample standard deviations. They also find that for this series length, over fitting the model can lead to increased sample standard deviations, although often does not significantly alter the results, whilst under fitting leads to smaller sample standard deviations with much

larger sample biases. They also consider times series of length $n = 100$ generated from Gaussian FARIMA(0, d_0 ,0) processes and find that, when the correct model is fitted, sample biases less than 0.05 in modulus and sample standard deviations less than 0.95 are reported, whilst over fitting for this series length can lead to greatly increases sample biases and sample standard deviations.

Bhansali and Kokoszka [2001a] also consider the effects of over and under fitting the model when the true model is a FARIMA(p,d,q) model with $p, q \in (0, 1, 2)$ and find for time series of length $n = 1000$ this can lead to increased sample biases and sample standard deviations. They then consider use of criteria to choose the model for each time series. In particular, they compare AIC and BIC, show that both can lead to reductions in bias and standard deviation compared to arbitrarily fitting a model, and find BIC to be preferred overall.

Following from these works, the simulation study presented here first investigates the sampling properties of the estimated parameters when the model is correctly specified. Results are then shown from when the model is over and under parameterised. Finally, the use of AIC and BIC to fit the model are considered.

The initial study focuses on the following four models

- 1) DFARIMA(0, $d = 0.25, c = 1, 0$)
- 2) DFARIMA(ar=0.5, $d = 0.25, c = 1, 0$)
- 3) DFARIMA(0, $d = 0.25, c = 1, ma=0.25$)
- 4) DFARIMA(ar=0.5, $d = 0.25, c = 1, ma=0.25$)

with Gaussian innovations. These are simulated using the method proposed by Davis and Harte [1987], see Theorem 5.4.1, from the spectral density given in equation 5.96. The ar and ma parameters were chosen before any simulations were carried out with the only consideration being that they not be equal to insure no common root between the ar and ma polynomials. For each model, three series lengths are considered, $n_1 = 100$, $n_2 = 1000$ and $n_3 = 10,000$. Since the asymptotic distribution of the parameter estimates, given by Theorem 2.3.1, shows the parameter estimates to be \sqrt{n} -consistent it was assumed a series length of $n = 10,000$ would be sufficient and thus a series length of $n = 10^5$ is not considered in this section. Each simulated time series is replicated 1000 times.

Taking the values of p and q to be known, for each simulated time series the estimates of the parameters were found by maximising the Whittle likelihood approximation. The sample biases, sample standard deviations and p-values for the Shapiro-Wilk test of normality with the null hypothesis of a normal distribution are given in table 5.8. Since p and q are known, Theorem 2.3.1 is applicable and the theoretical asymptotic standard deviations have been calculated using the discrete approximation method mentioned in section 5.3 with \tilde{n} set to 10^7 .

It can be seen in table 5.8 that both the biases and standard deviations decrease as the series length n increases and the p-values tend to increase as the series length n increases. This fits with the asymptotic theory of Theorem 2.3.1, which states the estimates are consistent and asymptotically normally distributed.

DFARIMA(0, $d = 0.25$, $c = 1, 0$)												
n	Sample Bias				Sample Standard Deviation				P-Value			
	d	c	ar	ma	d	c	ar	ma	d	c	ar	ma
100	-0.080	-0.058	-	-	0.204 (0.15)	0.230 (0.19)	-	-	0.0	0.1	-	-
1000	-0.014	-0.012	-	-	0.054 (0.05)	0.065 (0.06)	-	-	0.2	0.6	-	-
10,000	-0.004	-0.003	-	-	0.014 (0.01)	0.020 (0.02)	-	-	0.6	0.3	-	-
DFARIMA(ar=0.5, $d = 0.25$, $c = 1, 0$)												
n	Sample Bias				Sample Standard Deviation				P-Value			
	d	c	ar	ma	d	c	ar	ma	d	c	ar	ma
100	-0.210	-0.208	-0.039	-	0.286 (0.22)	0.359 (0.21)	0.367 (0.28)	-	0.0	0.0	0.0	-
1000	-0.061	-0.024	0.031	-	0.137 (0.07)	0.078 (0.07)	0.131 (0.09)	-	0.0	0.0	0.0	-
10,000	-0.004	0.000	0.005	-	0.024 (0.02)	0.027 (0.02)	0.032 (0.03)	-	0.6	0.1	0.1	-
DFARIMA(0, $d = 0.25$, $c = 1, ma=0.25$)												
n	Sample Bias				Sample Standard Deviation				P-Value			
	d	c	ar	ma	d	c	ar	ma	d	c	ar	ma
100	-0.023	-0.007	-	-0.011	0.217 (0.55)	0.380 (1.76)	-	0.301 (1.23)	0.0	0.0	-	0.0
1000	-0.013	-0.046	-	-0.039	0.087 (0.17)	0.251 (0.56)	-	0.178 (0.39)	0.0	0.0	-	0.0
10,000	-0.004	-0.016	-	-0.014	0.045 (0.05)	0.146 (0.18)	-	0.103 (0.12)	0.2	0.7	-	0.6
DFARIMA(ar=0.5, $d = 0.25$, $c = 1, ma=0.25$)												
n	Sample Bias				Sample Standard Deviation				P-Value			
	d	c	ar	ma	d	c	ar	ma	d	c	ar	ma
100	-0.190	-0.131	0.022	-0.003	0.293 (0.56)	0.368 (1.79)	0.321 (0.28)	0.254 (1.24)	0.0	0.0	0.0	0.0
1000	-0.072	-0.057	0.054	-0.047	0.129 (0.18)	0.240 (0.57)	0.114 (0.09)	0.185 (0.39)	0.0	0.0	0.0	0.0
10,000	-0.010	-0.024	0.010	-0.025	0.043 (0.06)	0.159 (0.18)	0.036 (0.03)	0.131 (0.12)	0.5	0.0	0.1	0.0

Tab. 5.8: Results of estimating the parameters of the DFARIMA models when the true model is fitted. theoretical standard deviations are shown in brackets beneath the sample standard deviations. theoretical biases are all zero. The p-values stated are for the Shapiro-Wilk test of normality with the null hypothesis of a normal distribution.

**ORIGINAL COPY TIGHTLY
BOUND**

The sample biases for all four models with series length $n \geq 1000$ are of modulus less than 0.072. For models 1 and 3 this is also true for the sample biases of the estimates with series length $n = 100$. For models 2 and 4, the biases of the estimates of d_0 and c_0 are larger with a modulus of around 0.2. However, even for these models the sample biases of the ar and ma coefficients are small for a series length of $n = 100$. The asymptotic theory that the parameter estimates are unbiased thus seems to hold reasonably well for series lengths of moderate size.

The p-values for the Shapiro-Wilk test of normality with the null hypothesis of a normal distribution, given in table 5.8, are correct to 1 dp. Thus a value of 0.0 represents the sampling distribution is significantly different from normality at the 5% significance level. It can be seen that for model 1, the sampling distributions were not significantly different from the normal distribution for series lengths of $n \geq 1000$, and the sampling distribution of the estimate of c_0 was not significantly different from the normal distribution even for a series length of $n = 100$. For models 2-4, the sampling distributions of the estimates did show significant different from the normal distribution for series lengths $n \leq 1000$, however for a series length of $n = 10,000$ the majority of the sampling distributions of the estimates from these models were not significantly different from the normal distribution. This reinforces the results of Theorem 2.3.1 which state that the estimates are asymptotically normally distributed.

Comparison of the theoretical asymptotic standard deviations, see Theorem 2.3.1, with the sample standard deviations, shown in table 5.8, indicates a close fit to the asymptotic results when $n = 10,000$ for all the estimates from all four models being considered. For series lengths $n \leq 1000$ the theoretical asymptotic standard deviations are often quite different from the sampling standard deviations reported, particularly for the estimates of c_0 and the ma coefficient in models 3 and 4. Interestingly, however, the difference is often that the sampling standard deviations are actually much smaller than the asymptotic theory would suggest, although the results for model 2 have sampling standard deviations slightly larger than the theoretical ones.

The results given in table 5.8 show the sample estimates generally seem to agree with the asymptotic results of Theorem 2.3.1 in that they have small biases and standard deviations and seem to tend to a normal distribution as the series length increases. However, these results are only applicable when the true values of p and q are known. Since this is rarely the case in applications, results are now presented assuming p and q to be unknown. In addition, it of interest to see some possible effects on estimating the parameters when $c_0 \neq 0$ is falsely assumed to be zero and also when $c_0 = 0$ is falsely assumed to be present.

In view of this, five further models are presented

- 5) FARIMA(0, $d = 0.25, 0$)
- 6) FARIMA(ar=0.5, $d = 0.25, 0$)
- 7) FARIMA(0, $d = 0.25, ma=0.25$)
- 8) FARIMA(ar=0.5, $d = 0.25, ma=0.25$)

Over Parameterised						
True Model: DFARIMA(0,d = 0.25, c = 1,0)						
	Fitted model: DFARIMA(1,d,c,1)			Fitted model: FARIMA(1,d,1)		
n	100	1000	10,000	100	1000	10,000
Sample Bias of \hat{d}	-0.216	-0.054	-0.011	-0.329	-0.293	-0.267
Sample SD of \hat{d}	0.327	0.163	0.050	0.250	0.058	0.015
True Model: FARIMA(0,d = 0.25,0)						
	Fitted model: DFARIMA(1,d,c,1)			Fitted model: FARIMA(1,d,1)		
n	100	1000	10,000	100	1000	10,000
Sample Bias of \hat{d}	-0.095	0.129	0.163	-0.199	-0.036	-0.006
Sample SD of \hat{d}	0.318	0.185	0.143	0.338	0.097	0.028
Under Parameterised						
True Model: DFARIMA(ar=0.5,d = 0.25, c = 1,ma=0.25)						
	Fitted model: DFARIMA(0,d,c,0)			Fitted model: FARIMA(0,d,0)		
n	100	1000	10,000	100	1000	10,000
Sample Bias of \hat{d}	0.034	0.043	0.040	-0.238	-0.204	-0.196
Sample SD of \hat{d}	0.205	0.057	0.016	0.103	0.028	0.009
True Model: FARIMA(ar=0.5,d = 0.25,ma=0.25)						
	Fitted model: DFARIMA(0,d,c,0)			Fitted model: FARIMA(0,d,0)		
n	100	1000	10,000	100	1000	10,000
Sample Bias of \hat{d}	0.017	0.039	0.036	0.670	0.630	0.611
Sample SD of \hat{d}	0.214	0.056	0.017	0.125	0.031	0.007

Tab. 5.9: Mean sample bias and standard deviations for the estimated values of d_0 when the fitted model has been over and under parameterised.

9) A sequence of IID $N(0,1)$

with Gaussian innovations. Models 5 - 8 are equivalent to models 1-4 with $c_0 = 0$. As previously, these are simulated using the method proposed by Davis and Harte [1987], see Theorem 5.4.1. Model 9, that is the sequence of IID $N(0,1)$, is included for completeness.

The results given in table 5.9 show the sample biases and standard deviations which occur when estimating d_0 whilst over or under fitting the model. In order to investigate the effects of over fitting the model, series are generated from models 1 and 5, which have $p, q = 0$, and the estimates of d_0 are estimated from fitting FARIMA(1,d,1) and DFARIMA(1,d,c,1) models. The investigation of the effects of under fitting reverses this, with series generated from models 4 and 8, which have $p, q = 1$, and estimates found fitting FARIMA(0,d,0) and DFARIMA(0,d,c,0) models.

When fitting a FARIMA model to the data, the results tend to agree with those presented in Taqqu and Teverovsky [1996]. When a FARIMA(1,d,1) model is fitted to a FARIMA(0,d,0) process, i.e. the model is overfitted, the bias seems to be small for $n \geq 1000$ and the sample standard deviations, although reasonably small, are large compared to the sample standard deviations shown when a FARIMA(0,d,0) model is fitted. When a FARIMA(0,d,0) model is fitted to a FARIMA(1,d,1) process, i.e. the model is underfitted, the bias is much larger with smaller standard deviations. Similar results for under fitting also appear when fitting a FARIMA(0,d,0) model to a DFARIMA(1,d,c,1).

The case of fitting a FARIMA(1,d,1) model to a DFARIMA(0,d,c,0) process behaves similar to the underfitting cases, resulting in large bias and moderately small standard deviations. Indeed, the fitting of a FARIMA(1,d,1) model to a DFARIMA(0,d,c,0) process, although overfitting the values of p and q , essentially underfits the presence of c_0 . This suggests increasing the number of short range parameters does not compensate for ignoring the presence of c_0 .

The results from fitting a DFARIMA(p,d,c,q) model, also shown in table 5.9, seem somewhat different. The results from fitting a DFARIMA(1,d,c,1) model to a DFARIMA(0,d,c,0) process, an example of overfitting, seem to follow the general pattern in that the biases are small, for $n \geq 1000$, whilst the standard deviations are larger than those reported in table 5.8 when the true model was fitted. However, the results when a DFARIMA(1,d,c,1) model is fitted to a FARIMA(0,d,0) process, another example of overfitting, are very different. Both biases and standard deviations seem much larger. In addition, although the standard deviations appear to decrease as the series length n increases, the mean bias actually appears to be increasing.

The results for under fitting a DFARIMA(0,d,c,0) model to DFARIMA(1,d,c,1) and FARIMA(1,d,1) processes also vary from those for underfitting a FARIMA model. They appear to give both small biases and small standard deviations. The

parameter c thus seems to be able to compensate for the underfitting of p and q . It may be of interest for future research to examine how well this result holds for larger values of p and q and how dependent it is on the choice of the short range parameters.

DFARIMA(0, $d = 0.25, c = 1, 0$)									
n	Model fitted using AIC				Model fitted using BIC				
	Sample Bias		Sample SD		Sample Bias		Sample SD		
	d	c	d	c	d	c	d	c	
100	-0.254	-0.551	0.262	0.525	-0.269	-0.553	0.276	0.525	
1000	-0.104	-0.274	0.165	0.447	-0.099	-0.255	0.161	0.435	
10,000	-0.020	-0.060	0.067	0.247	-0.004	-0.003	0.014	0.020	
DFARIMA(ar=0.5, $d = 0.25, c = 1, 0$)									
n	Model fitted using AIC				Model fitted using BIC				
	Sample Bias		Sample SD		Sample Bias		Sample SD		
	d	c	d	c	d	c	d	c	
100	-0.103	-0.843	0.204	0.078	-0.178	-0.959	0.190	0.174	
1000	-0.097	-0.226	0.130	0.403	-0.080	-0.476	0.111	0.466	
10,000	-0.001	0.012	0.031	0.078	-0.004	0.000	0.024	0.027	
DFARIMA(0, $d = 0.25, c = 1, ma = 0.25$)									
n	Model fitted using AIC				Model fitted using BIC				
	Sample Bias		Sample SD		Sample Bias		Sample SD		
	d	c	d	c	d	c	d	c	
100	-0.300	-0.765	0.232	0.338	-0.387	-0.830	0.280	0.325	
1000	-0.175	-0.565	0.121	0.354	-0.162	-0.573	0.103	0.321	
10,000	-0.019	-0.060	0.057	0.208	-0.081	-0.321	0.033	0.118	
DFARIMA(ar=0.5, $d = 0.25, c = 1, ma = 0.25$)									
n	Model fitted using AIC				Model fitted using BIC				
	Sample Bias		Sample SD		Sample Bias		Sample SD		
	d	c	d	c	d	c	d	c	
100	-0.174	-0.938	0.171	0.210	-0.239	-0.999	0.075	0.022	
1000	-0.134	-0.391	0.101	0.220	-0.139	-0.516	0.117	0.296	
10,000	-0.021	-0.064	0.054	0.205	-0.082	-0.315	0.024	0.020	

Tab. 5.10: Mean sample bias and standard deviations for the estimated values of d_0 and c_0 when the true model is assumed unknown and a model is fitted using either AIC or BIC from a choice of DFARIMA and FARIMA models.

In the study presented here, having considered some possible effects of over and under fitting, attention now turns to model fitting. The work of Beran et al [1998] show the suitability of using AIC and BIC for fitting FAR(p) models and suggest a wider suitability to the fitting of other long memory models. They show BIC generally gives consistent estimates of p whilst AIC overestimates p around 30% of the time.

Time series were generated from models 1-4 as before. The true model was then assumed unknown and AIC and BIC were then used to choose from models 1-9. Table 5.10 shows the mean sample bias and standard deviations for the estimated values of d_0 and c_0 . Note, when the model fitted was a FARIMA(p,d,q) model the value of \hat{c} was taken to be zero.

Comparison with the sample biases and standard deviations from the estimates of d_0 and c_0 when the true models are fitted, given in table 5.8, show the biases tend to be larger, whilst the standard deviations for model 1 seem to be larger, the standard deviations for models 2 and 3 seem similar in magnitude and the standard deviations for model 4 seem generally smaller.

For all four models the biases and standard deviations are seen to decrease as n increases when using either AIC or BIC in model fitting. The large biases and standard deviations for series length $n = 100$ suggest this is too short. The results for $n \geq 1000$ are much more acceptable.

In comparing the results using AIC with those using BIC, the results for AIC seem to have smaller mean biases the majority for the majority of the results, whilst the results for BIC seem to have reduced standard deviations.

The proportion of times each model was chosen using AIC and BIC for series length $n = 10,000$ are shown in figure 5.3. It can be seen when the true model is either 1 or 2, both AIC and BIC correctly identify this model the majority of times, with BIC slightly outperforming AIC. When the true model is either 3 or 4 though, it can be seen that AIC correctly identifies the model around 50% of the time, whilst BIC tends to select models 1 and 2 respectively. In view of the results for underfitting given in table 5.9 this is perhaps not surprising, since the underfitted model still seems to give a good fit.

Also included in figure 5.3 are the proportions of each model chosen when the series were generated from models 5-9, the cases when $c_0 = 0$. For these models, both AIC and BIC correctly identify the model the majority of the time, with BIC outperforming AIC. This is consistent with previous studies, see for example Beran et al [1998]. Particularly worth noting is that both AIC and BIC correctly seem to distinguish between cases when $c_0 = 0$ and $c_0 \neq 0$ the vast majority of the times. The large increase in both bias and standard deviation seen in table 5.9 when DFARIMA(1,d,c,1) models were fitted to FARIMA(0,d,0) processes should thus not be too serious an issue in practise when the model is fitted using a suitable criterion.

Finally, table 5.9 has previously shown that arbitrarily fitting a FARIMA(p,d,q) process can lead to large biases. However, since the majority of the literature to date has ignored the possibility of $c_0 \neq 0$, it is of interest to explore further how the fitting of FARIMA(p,d,q) models to DFARIMA(p,d,c,q) processes may effect the estimates of d_0 .

Table 5.11 shows the biases and standard deviations for the estimates of d_0 when the true model, assumed unknown, is one of models 1-4, whilst the model fitted using either AIC or BIC is restricted in choice to FARIMA(p,d,q) models with $p, q \in (0, 1)$.

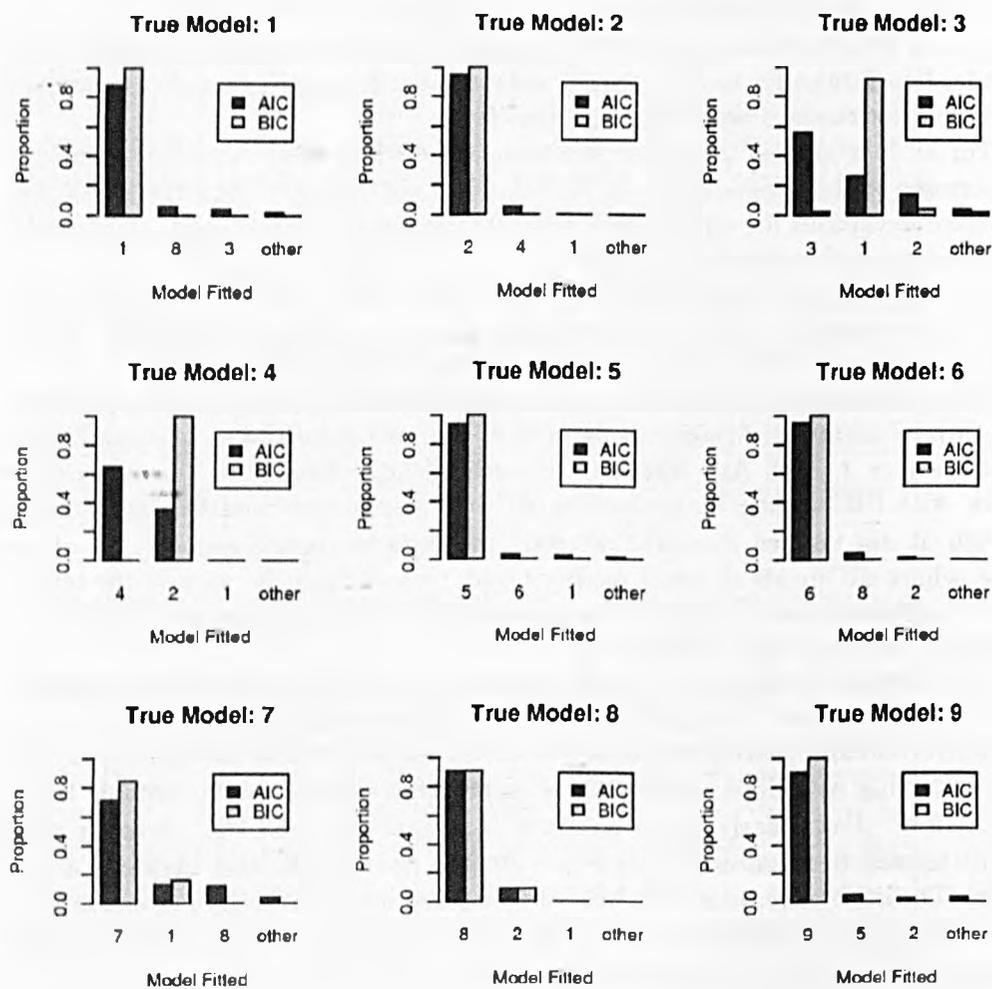


Fig. 5.3: Models chosen by AIC and BIC when series length $n = 10,000$.

Model fitted using AIC								
Sample Bias					Sample SD			
True Model	1	2	3	4	1	2	3	4
n								
100	-0.355	-0.133	-0.346	-0.202	0.232	0.164	0.209	0.273
1000	-0.296	-0.145	-0.308	-0.091	0.060	0.040	0.052	0.127
10,000	-0.267	-0.141	-0.289	-0.062	0.015	0.012	0.017	0.095
Model fitted using BIC								
Sample Bias					Sample SD			
True Model	1	2	3	4	1	2	3	4
n								
100	-0.389	-0.115	-0.354	-0.153	0.200	0.165	0.196	0.248
1000	-0.310	-0.142	-0.301	-0.086	0.068	0.044	0.047	0.118
10,000	-0.267	-0.141	-0.285	-0.061	0.015	0.012	0.018	0.093

Tab. 5.11: Mean sample bias and standard deviations for the estimated values of d_0 when the true model is assumed unknown and a model is fitted using either AIC or BIC from a choice of FARIMA models.

It can be seen that even for series length $n = 10,000$ the sample biases using both AIC and BIC to fit the model can be substantial. Indeed, the biases for models 1 and 3 using both AIC and BIC are greater in modulus than d_0 and would thus suggest intermediate memory when in fact quite strong long memory is present.

In summary, when the correct model is known, the empirical results give weight to the results of Theorem 2.3.1 which state the estimates of the parameters are asymptotically \sqrt{n} -consistent and normally distributed. When the correct model is unknown, it has been found that use of AIC and BIC can distinguish between the cases when $c = 0$ and $c \neq 0$ and tend to fit models which lead to good estimates of d_0 . Finally, since restricting the range of models available to either $c = 0$ or $c \neq 0$ can lead to large biases, it is strongly recommended that a range of FARIMA(p,d,q) and DFARIMA(p,d,c,q) models be taken into account when fitting a model.

5.4.3 Application to Other Series

In this section, the new methods of estimating d and c are applied to the processes given by Martin and Eccleston [1992], Martin and Walker [1997] and Palma [2007] and the further example of 'weak' long memory due to the unpublished work of Prof R. Bhansali. Application to the much studied Bellcore data is also then made. Previous studies of data from the same source include Clegg [2006] and Leland and Wilson [1991] and Leland et al [1993].

The three processes defined by equations 5.3, 5.4 and 5.5 will be referred to as processes I, II and III respectively. Process I, that of Martin and Eccleston [1992], Martin and Walker [1997], is defined with $A = 0.5$ for this study. For each of these

processes 1000 series of length 10^5 were generated using the method of Theorem 5.4.1, with Gaussian innovations. The new dual local Whittle and DGPH methods were then applied with $m = 1000$ to find estimates of d and c . The standard local Whittle and GPH methods were also applied with $m = 1000$ for comparison and to see if the theoretical biases given by Theorems 5.2.1 and 5.1.3 respectively appear to hold.

A DFARIMA approach to estimation was also applied, but due to higher requirements of computational time this method was only applied to the first 10^4 observations from each series. The approach is to fit DFARIMA(p,d,c,q) and FARIMA(p,d,q) models to the data using the Whittle likelihood estimates of the parameters with $p, q \in (0, 1, 2)$ and choose from these fitted models using a suitable criterion. The use of AIC and BIC were applied here.

Figures 5.4, 5.5 and 5.6 show the Normal Q-Q plots of the estimates of d for the series generated from processes I, II and III, respectively. It can be seen that the distributions of the new dual local Whittle and DGPH estimates of d appear very close to normality for all three processes. The standard local Whittle and GPH estimates also appear close to normality for all three processes, but appear to deviate slightly further than the new dual parameter estimates near the tails.

The estimates of the DFARIMA method for process I are significantly non-Gaussian using both AIC and BIC, although when using BIC the distribution of the majority of the estimates appears well approximated by the normal distribution, but the overall distribution appears bimodal. For process II, the estimates from the DFARIMA method using AIC appear to have very heavy tails but the estimates using BIC appear much closer to normality. For process III the estimates from the DFARIMA method using both AIC and BIC appear reasonably close to normality, with some deviate in the upper tail behaviour.

Since the estimate of c is often strongly correlated with that of d , the Normal Q-Q plots of the estimated values of c are similar in shape to those shown for d and thus omitted.

For processes I and II the models chosen by AIC and BIC were DFARIMA models with $c \neq 0$ for 100% of the simulated series. For process III however, only 1.3% of the fitted models using both AIC and BIC were DFARIMA models with $c \neq 0$. The remaining 98.7% of the models fitted were FARIMA models.

The sample means and standard deviations of the estimates of both d and c are shown in table 5.12. The theoretical values of the bias and standard deviations for the standard local Whittle and GPH methods and the standard deviations for the new dual Local Whittle and DGPH methods are calculated as in section 5.4.1. For the DFARIMA method, since no 'true' finite parameter DFARIMA model exists for these processes, the theoretical standard deviations for each fitted model were calculated as in section 5.4.2 and a 90% coverage of these theoretical standard deviations, removing the upper and lower 5%, is reported. For process III it can be seen that the theoretical standard deviation of c is reported as 0. This is due to less than 5%

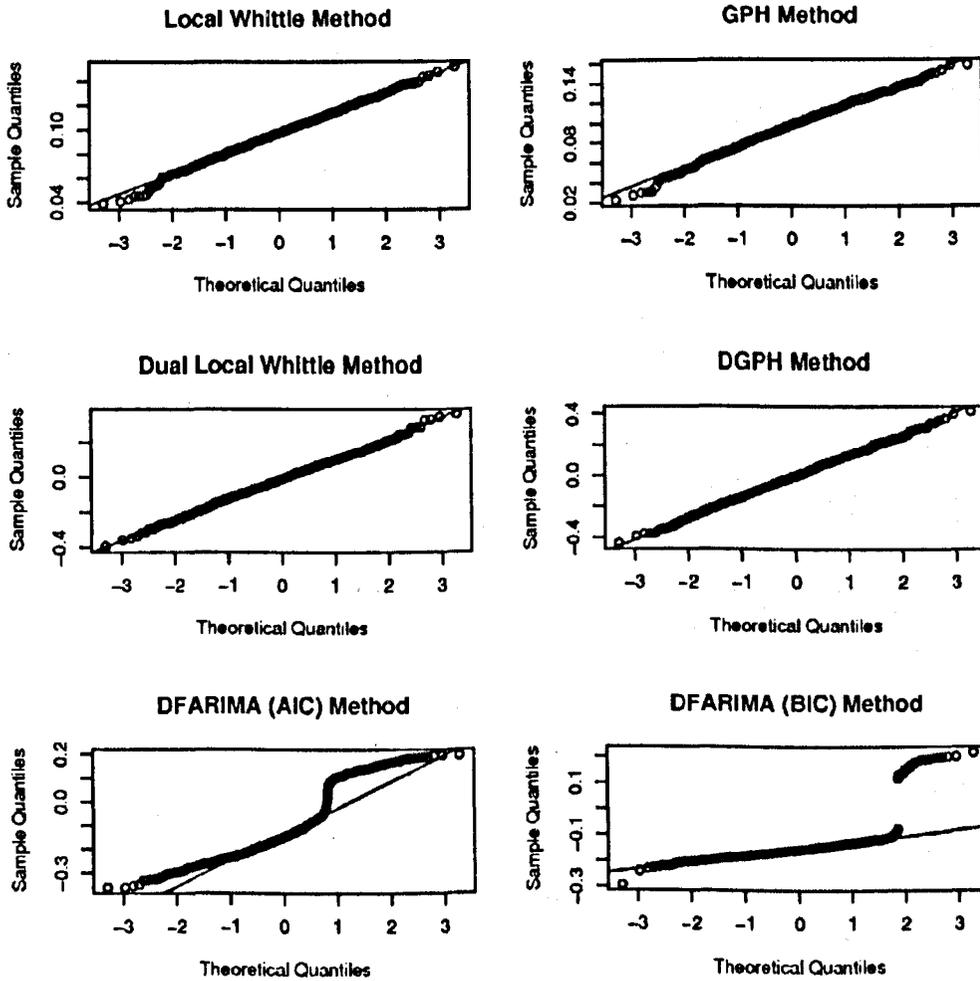


Fig. 5.4: Normal Q-Q plots of the estimates of d for the series generated from process I.

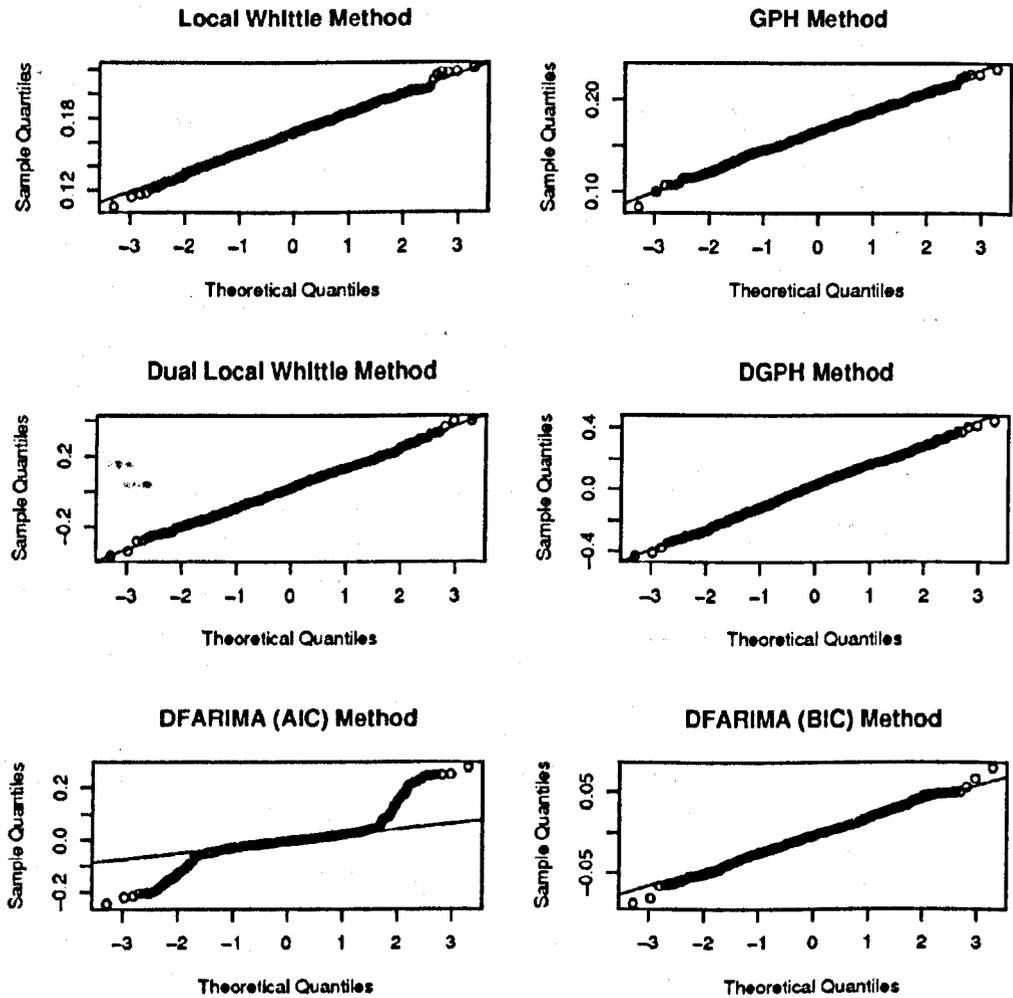


Fig. 5.5: Normal Q-Q plots of the estimates of d for the series generated from process II.

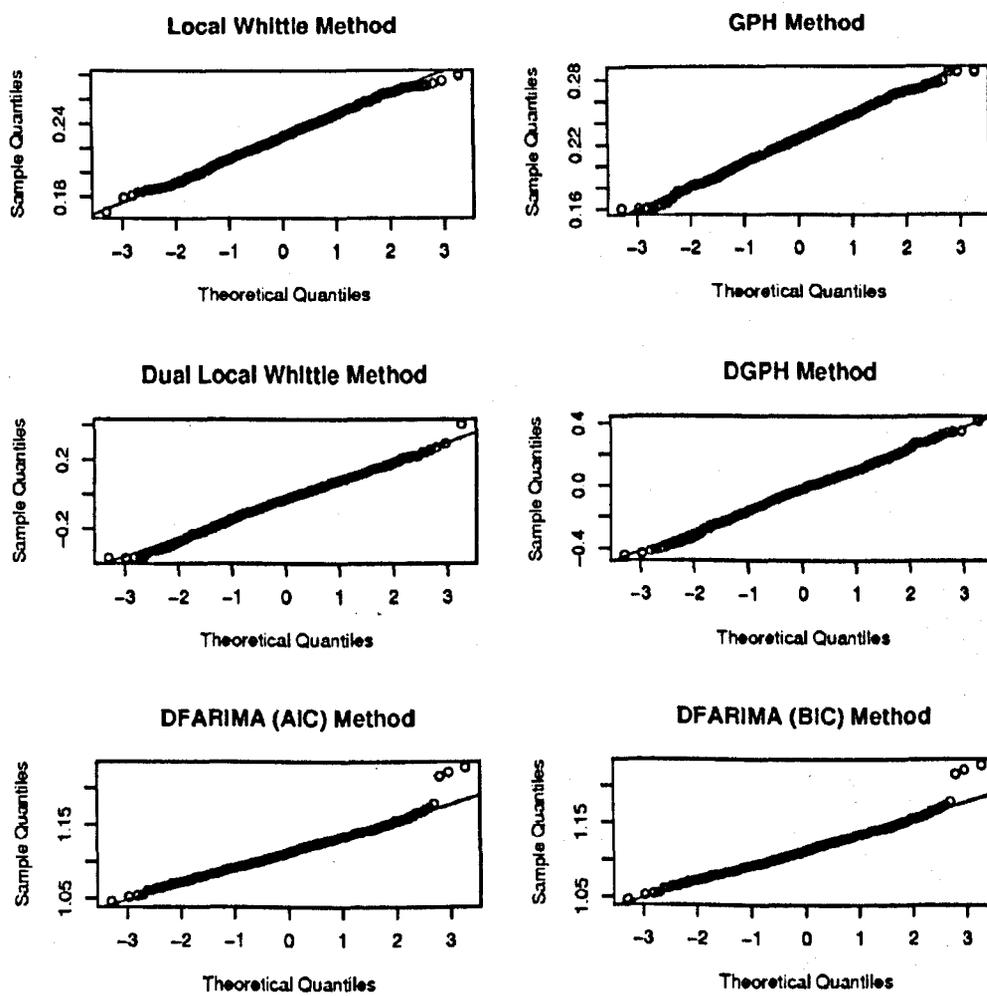


Fig. 5.6: Normal Q-Q plots of the estimates of d for the series generated from process III.

Estimate	Series	Sample			theoretical		
		I	II	III	I	II	III
Local Whittle	mean	0.097	0.168	0.229	0.13	0.26	0.26
	sd	0.017	0.017	0.018	0.016	0.016	0.016
GPH	mean	0.097	0.166	0.226	0.11	0.23	0.23
	sd	0.021	0.021	0.022	0.02	0.02	0.02
DLW d	mean	-0.011	0.021	-0.027	0	0	0
	sd	0.114	0.113	0.107	0.091	0.091	0.091
DLW c	mean	-0.469	-0.636	-1.106	-0.5	-1	-1
	sd	0.495	0.486	0.459	1.048	1.048	1.048
DGP1I d	mean	-0.005	0.025	-0.022	0	0	0
	sd	0.136	0.137	0.132	0.127	0.127	0.127
DGP1I c	mean	-0.445	-0.619	-1.089	-0.5	-1	-1
	sd	0.591	0.591	0.568	0.550	0.550	0.550
DFARIMA (AIC) d	mean	-0.109	-0.003	1.114	0	0	0
	sd	0.135	0.052	0.023	(0.02,0.09)	(0.02,0.04)	(0.02,0.04)
DFARIMA (AIC) c	mean	-1.049	-0.995	0.006	-0.5	-1	-1
	sd	0.443	0.168	0.056	(0.03,0.21)	(0.03,0.16)	0
DFARIMA (BIC) d	mean	-0.148	-0.004	1.114	0	0	0
	sd	0.062	0.022	0.023	0.02	0.02	(0.02,0.04)
DFARIMA (BIC) c	mean	-1.133	-1.003	0.006	-0.5	-1	-1
	sd	0.173	0.027	0.056	0.03	0.03	0

Tab. 5.12: Means and standard deviations for the estimated values of d and c . The values in brackets represent a 90% coverage.

of the models chosen containing c . For the DFARIMA method using BIC for the other two processes the theoretical standard deviations are reported as a constant since over 95% of the simulations had a DFARIMA(0,d,c,0) model fitted and the theoretical standard deviations do not depend on c or d .

For the standard local Whittle and GPII estimates, the sample standard deviations appear to agree with the asymptotic theory, whilst the biases appear slightly smaller than expected, particularly for process II. As with the results presented in section 5.4.1, the sample standard deviations of the dual local Whittle estimates of d are larger than the asymptotic theory suggest, whilst the sample standard deviations of the estimates of c are smaller. The sample standard deviations of the estimates of d and c once again agree much more with the theory. The means of the estimates of d for the dual local Whittle and DGPII methods are close to zero for all three processes. The means of the estimates of c for the dual local Whittle and DGPII methods are close to -0.5 and -1 for processes I and III respectively, in agreement with the asymptotic theory. The means of the estimates of c for the dual local Whittle and DGPII methods for process II however are around -0.6 , whilst the asymptotic theory suggests $c = -1$. The smaller than expected biases of the standard local Whittle and GPII estimates for process II also suggest a value of c such that $-1 < c < -0.5$.

The results for the DFARIMA method for process II however give estimates of d and c close to 0 and -1 respectively and thus support the asymptotic theory. The standard deviations are of similar magnitude to those suggested by the asymptotic theory, though appear larger when using AIC. The results of the DFARIMA method for process I show the means of the estimates of d and c are negatively biased using both AIC and BIC, whilst the sample standard deviations are generally much larger than allowed for by the theory. For process III, the majority of models fitted were FARIMA(p,d,q) models and thus the mean of c is close to zero. The mean of d for using both AIC and BIC can be seen to be very strongly biased, suggesting instead of 'weak' long memory that the process has a unit root. This bias is, however, not due to the fitting of FARIMA instead of DFARIMA models, as the fitted DFARIMA models also had unstationary values of d . The cause of the bias is likely caused by the behaviour of the spectral density away from zero, in particular the fact that $f(\pi/3) = 0$, a situation which is not compatible with the DFARIMA and FARIMA spectral densities.

Results are now presented for the Bellcore Ethernet data. Previous studies of data from the same source can be seen in Clegg [2006], Leland and Wilson [1991] and Leland et al [1993], however the results reported here are for a different time period. The data gives the time and size of packet arrivals seen on an Ethernet at the Bellcore Morristown Research and Engineering facility during October 1989. From these, two series are generated. The interarrival times series records the times between packet arrivals and the bytes/0.01s series records the total amount of bytes received for each 0.01 second time period. The data contains the information of

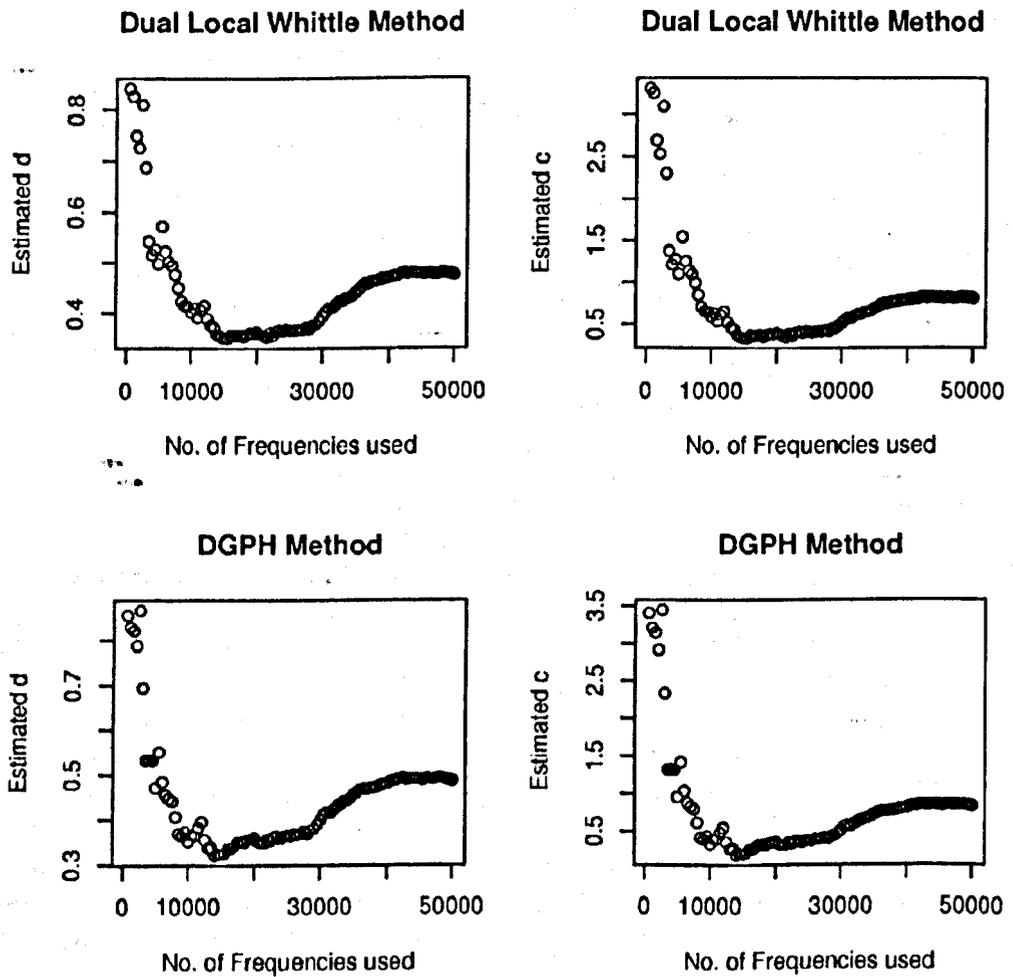


Fig. 5.7: Estimates of d and c for the Bellcore interarrival times series using the dual local Whittle and DGPH methods for various values of m .

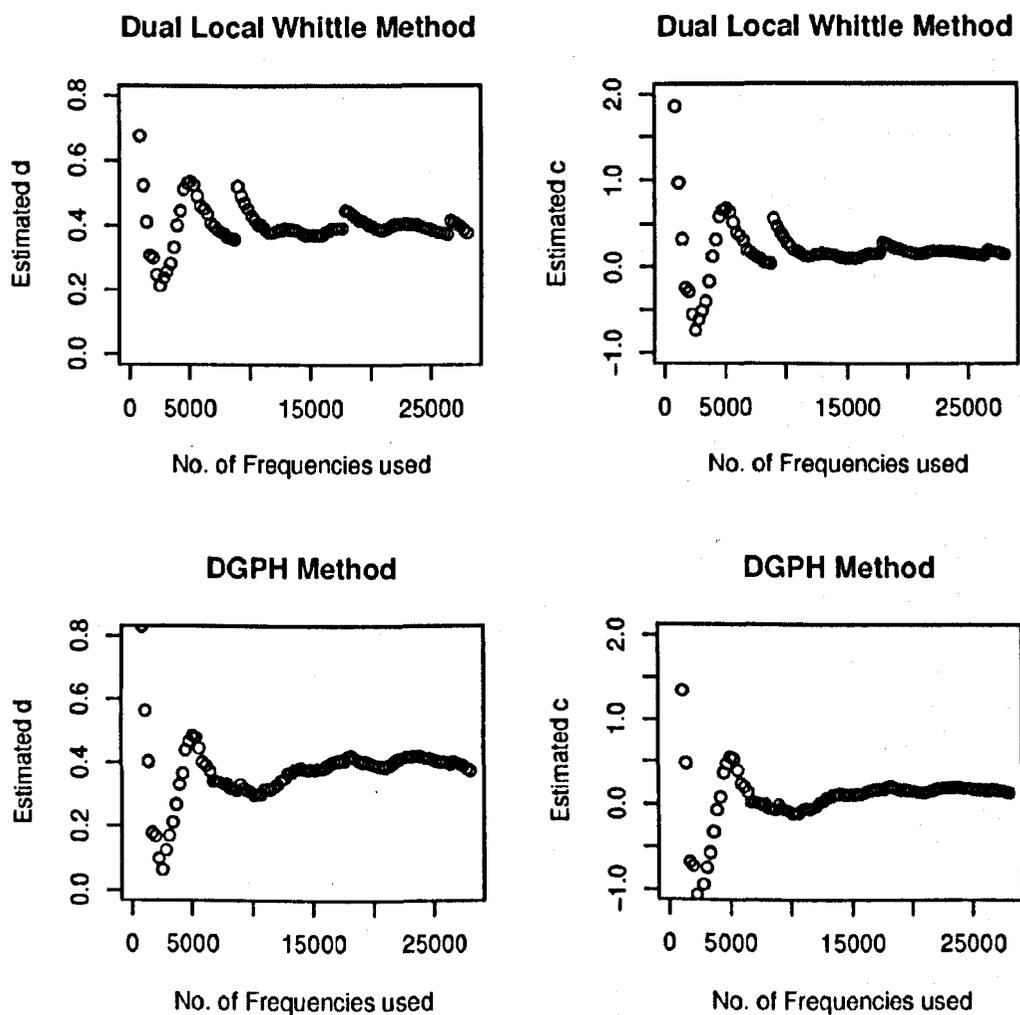


Fig. 5.8: Estimates of d and c for the Bellcore bytes/0.01s series using the dual local Whittle and DGP methods for various values of m .

1,000,000 packet arrivals over 1759.62 seconds. The interarrival times series is thus length 999,999 and the bytes/0.01s series is length 175,962. Similar series studied by Clegg [2006] suggest d approximately in (0.155, 0.246) for the interarrival times series and d approximately in (0.12, 0.325) for the bytes/0.01s series when c is assumed to be zero.

The DFARIMA method was applied to the first 10^4 observations of each series. For the interarrival times series, the use of AIC and BIC both agreed on a DFARIMA(1,d,c,0) model being fitted, with $d=0.439$, $c=0.926$ and $\phi=0.505$. For the bytes/0.01s series, the use of AIC and BIC again agreed on a DFARIMA(1,d,c,0) model being fitted, with similar parameters of $d=0.469$, $c=0.865$ and $\phi=0.646$. This suggests d may actually be larger than previously thought for these series.

The dual local Whittle and DGPII methods were applied to the full length of each series. Plots of the estimates of d and c using the dual local Whittle and DGPII methods for various values of m for the interarrival times series and the bytes/0.01s series can be seen in figures 5.7 and 5.8 respectively. For the interarrival times series, the estimates of d and c seem to be quite level for $m \in (15000, 27500)$ and also for $m \in (35000, 50000)$. For $m = 20,000$, the dual local Whittle estimates of d and c are 0.36 and 0.39 whilst the DGPII estimates of d and c are 0.36 and 0.36. For $m = 45,000$, the dual local Whittle estimates of d and c are 0.48 and 0.82 whilst the DGPII estimates of d and c are 0.49 and 0.85. Note, the results for $m \in (35000, 50000)$ are similar in magnitude as the estimates found using the DFARIMA method. The estimates for $m \in (15000, 275000)$ are smaller in size but still suggest d may be larger than previously thought.

For the bytes/0.01s series, the estimates of d and c seem to be quite level for $m \in (6000, 28000)$. For $m = 15,000$, the dual local Whittle estimates of d and c are 0.37 and 0.11 whilst the DGPII estimates of d and c are 0.38 and 0.12. These are similar in magnitude to the estimates for $m \in (15000, 275000)$ for the interarrival times series, again suggesting d may be larger than previously thought.

5.4.4 Application to Chaotic Intermittency Maps

The simulation study on the chaotic intermittency maps carried out in section 4.1 showed the standard methods of estimating d appeared strongly biased for the symmetric and asymmetric cusp maps, the logarithmic maps and the polynomial map with $\alpha \leq 0.5$. The asymptotic biases expected from the standard local Whittle and GPII methods are given by Theorems 5.2.1 and 5.1.3 respectively and these can now be compared to the empirical results for the chaotic intermittency maps.

In addition, the simulation study of the chaotic intermittency maps described in 4.1 was repeated and estimates of d and c were found using the new dual local Whittle, DGPII and DFARIMA methods, that is 1000 orbits of length 10^4 after a burn in time of $M = 10^7 - 10^4$ were retained for each map and the estimates of d

Map	c	Local Whittle		GPH		
		Asymptotic	Empirical	Asymptotic	theoretical	Empirical
Poly (0.3)	0	0	0.69	0	0	0.69
Poly (0.45)	0	0	0.20	0	0	0.19
Poly (0.5)	-0.5	0.11	0.12	0.36	0.12	0.10
Poly (0.65)	0	0	-0.01	0	0	-0.02
Poly (0.8)	0	0	-0.10	0	0	-0.10
Poly (0.9)	0	0	-0.03	0	0	-0.02
Symmetric	-0.5	0.11	0.28	0.36	0.12	0.24
Asymmetric	-0.5	0.11	0.31	0.36	0.12	0.26
Log (0.05)	0.525	-0.12	-0.21	-0.38	-0.12	-0.21
Log (0.15)	0.575	-0.13	-0.24	-0.42	-0.14	-0.24
Log (0.25)	0.625	-0.14	-0.27	-0.45	-0.15	-0.26
Log (0.3)	0.65	-0.15	-0.28	-0.47	-0.15	-0.28

Tab. 5.13: List of biases for the standard local Whittle and GPH estimates. The asymptotic values come from Theorems 5.2.1 and 5.1.3 respectively. The theoretical biases for the GPH method are found directly from equation 5.44. The empirical biases are the mean biases of those observed during the simulation study presented in section 4.1.

and c recorded for each. The initial values of each orbit were generated as described in section 4.1.1. To be comparable with the results from the standard estimation methods, the value for m for the dual local Whittle method was taken as 155 and the value of m for the DGPH was taken as 100.

For the DFARIMA method, the Whittle likelihood estimates were found for all DFARIMA(p, d, c, q) models with $p, q \in (0, \dots, 5)$ and the choice of model was obtained by minimising both the AIC and BIC. The fitting of FARIMA models was not included again as these results have previously been reported and generally seemed to fail to give a good fit.

Table 5.13 gives a list of sample biases for the standard local Whittle and GPH estimates taken from section 4.1 and the corresponding theoretical biases. The asymptotic values come from Theorems 5.2.1 and 5.1.3, whilst the theoretical biases for the standard GPH method are found directly from evaluating equation 5.44.

For the polynomial map with $\alpha < 0$, the estimated values of d show large positive bias, but since the asymptotic value of c for these maps is zero, this bias is not explained by Theorems 5.2.1 and 5.1.3. As previously mentioned in section 4.3, it is believed that if these maps converge to their asymptotic properties they do so at a very slow rate.

For the polynomial map with $\alpha = 0.5$, the asymptotic theory suggests the orbits possess long memory parameters of $d = 0$ and $c = -0.5$. It can be seen that the theoretical biases of the standard local Whittle and GPH estimates are quite close to the empirical results and thus this 'weak' long memory behaviour accounts for the biases seen. Also in agreement with the theoretical results, for the polynomial

map with $\alpha > 0.5$, the biases were small to begin with although some negative bias is seen for the polynomial map with $\alpha = 0.8$.

For the symmetric and asymmetric cusp maps, the asymptotic theory again suggests long memory parameters of $d = 0$ and $c = -0.5$. However, the theoretical biases seem to account for less than half of the bias seen in the empirical results. As discussed in section 4.3, the remainder of the bias is believed to be attributed to a slow convergence rate of the maps to their asymptotic properties and large autocorrelation at smaller lags. The results of the systematic sampling study shown in section 4.2.2 show the biases for the standard local Whittle and GPH estimates when sampling every 10th observation appear close to the theoretical biases of 0.11 and 0.12.

For the logarithmic maps the theoretical biases seem to account for roughly half of the bias of the empirical results. It can be seen that as β increases the value of c and thus the theoretical biases also increase. This dependence on β is reflected in the empirical results, although the rate of change of the sample biases with β seems faster than that suggested by the theoretical work.

Figure 5.9 shows the Normal Q-Q plots for the new dual parameter local Whittle estimates of d . It can be seen that other than in the tails, the distributions of the estimates appear close to normality. As with the distributions of the standard local Whittle estimates, the estimates from the logarithmic maps seem to deviate further from normality than for the other chaotic intermittency maps. The estimates of d from the DGPH method have quite similar distributions and the estimates of c for both methods are similarly distributed as the estimates of d due to strong correlation between the estimates.

The Normal Q-Q plots for the DFARIMA estimates of d using AIC and BIC are shown in figures 5.10 and 5.11 respectively. Similar to the results for the FARIMA estimates given in section 4.1, the distributions of the estimates of d from the DFARIMA methods appear to deviate further from normality than the estimates from the dual local Whittle and DGPH methods. Also the estimates found using AIC seem to be less normally distributed than those found using BIC. Comparison, however, with the Normal Q-Q plots for the FARIMA methods shows the DFARIMA estimates appear closer to normality and although heavy tailed behaviour is still apparent, the DFARIMA distributions appear unimodal.

The sample means and sample deviations for the dual local Whittle, DGPH and DFARIMA estimates of d and c are shown in table 5.14. For the polynomial map with $\alpha < 0.5$, the sample means of the dual local Whittle and DGPH estimates of d are less than zero showing signs of the intermediate memory suggested by the asymptotic theory. The means for the polynomial map with $\alpha = 0.45$ in particular seem very close to the asymptotic value of $d = -0.11$. However, the sample means of the dual local Whittle and DGPH estimates of both d and c show large negative bias for all other maps. The sample standard deviations of the dual local Whittle and DGPH estimates of both d and c can be seen to be unacceptably large for all the maps, making it difficult to draw meaningful conclusions from these results.

For all the maps, the sample means of the estimates of d and c from the DFARIMA

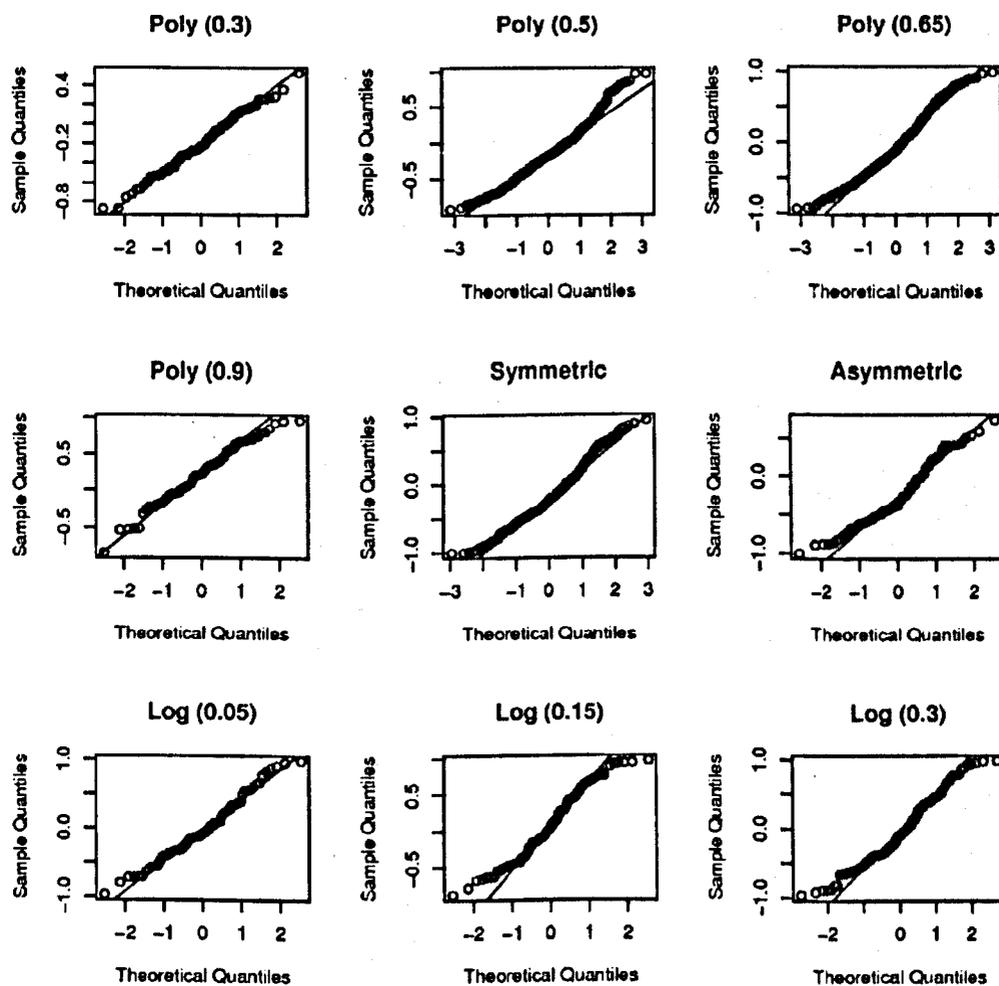


Fig. 5.9: Normal Q-Q plots for the new dual parameter local Whittle estimates of d .

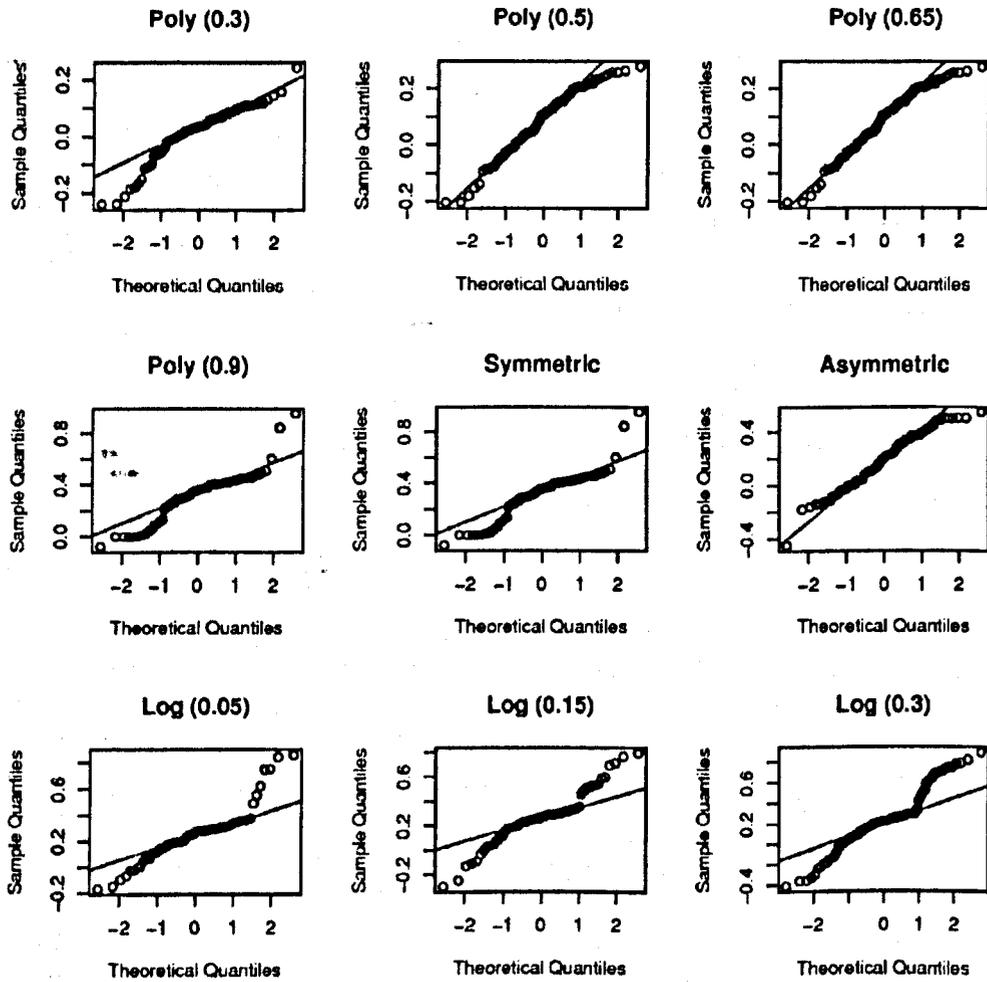


Fig. 5.10: Normal Q-Q plots for the DFARIMA estimates of d using AIC.

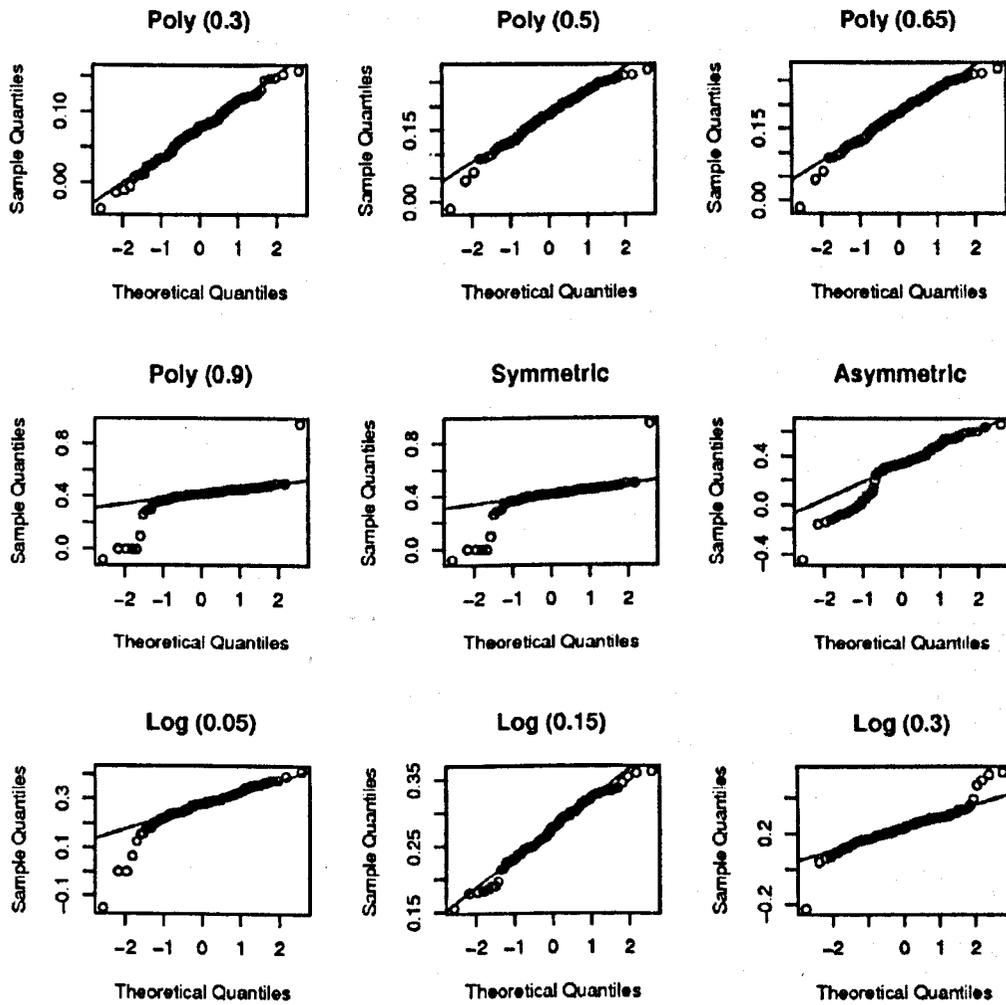


Fig. 5.11: Normal Q-Q plots for the DFARIMA estimates of d using BIC.

Sample Means										
Map	Asymptotic d	Asymptotic c	Dual Local Whittle		DGPH		DFARIMA (AIC)		DFARIMA (BIC)	
			d	\hat{c}	d	\hat{c}	d	\hat{c}	d	\hat{c}
Poly (0.3)	-0.67	0	-0.20	-0.84	-0.17	-0.75	0.024	-0.048	0.075	0.041
Poly (0.45)	-0.11	0	-0.13	-0.86	-0.10	-0.78	0.090	-0.12	0.18	0.062
Poly (0.5)	0	-0.5	-0.16	-1.0	-0.14	-0.96	0.092	-0.17	0.20	0.05
Poly (0.65)	0.23	0	-0.029	-0.95	-0.0010	-0.87	0.20	-0.18	0.30	0.063
Poly (0.8)	0.38	0	0.17	-0.92	0.10	-0.82	0.27	-0.34	0.35	-0.10
Poly (0.9)	0.44	0	0.32	-1.2	0.20	-0.57	0.29	-0.42	0.37	-0.17
Symmetric	0	-0.5	-0.19	-1.7	-0.18	-1.7	0.15	-0.85	0.42	0.019
Asymmetric	0	-0.5	-0.27	-2.2	-0.25	-2.1	0.20	-0.27	0.31	-0.12
Log (0.05)	0.5	0.525	0.062	-1.2	-0.04	-1.1	0.25	-0.053	0.27	-0.017
Log (0.15)	0.5	0.575	0.13	-0.62	0.030	-0.83	0.26	0.084	0.27	0.05
Log (0.25)	0.5	0.625	0.01	-0.83	-0.10	-1.0	0.19	0.0043	0.23	0.043
Log (0.3)	0.5	0.65	0.048	-0.70	-0.070	-1.0	0.22	0.022	0.23	0.051
Sample Standard deviations										
Poly (0.3)			0.30	1.1	0.35	1.3	0.084	0.28	0.037	0.036
Poly (0.45)			0.33	1.1	0.40	1.4	0.10	0.33	0.050	0.029
Poly (0.5)			0.36	1.2	0.41	1.4	0.15	0.50	0.075	0.13
Poly (0.65)			0.45	1.5	0.52	1.8	0.15	0.51	0.063	0.11
Poly (0.8)			0.45	2.0	0.45	1.5	0.19	1.1	0.16	1.0
Poly (0.9)			0.46	2.7	0.42	1.5	0.28	1.3	0.27	1.3
Symmetric			0.44	1.4	0.48	1.6	0.20	0.78	0.084	0.20
Asymmetric			0.42	1.4	0.48	1.6	0.20	0.59	0.21	0.48
Log (0.05)			0.51	2.2	0.47	1.5	0.16	0.66	0.078	0.49
Log (0.15)			0.53	1.9	0.49	1.6	0.18	0.56	0.044	0.055
Log (0.25)			0.54	1.9	0.42	1.5	0.22	0.71	0.070	0.12
Log (0.3)			0.57	2.0	0.52	1.7	0.24	0.79	0.075	0.17

Tab. 5.14: Sample means and Sample standard deviations for the new dual parameter estimates of d and c .

method using AIC are smaller than the corresponding sample means using BIC, whilst the sample standard deviations are smaller for the estimates found using BIC. For the DFARIMA method using both AIC and BIC the sample means of the estimates of d for the polynomial map with $\alpha < 0.5$ are once again positive and generally close to zero.

For the polynomial map with $\alpha = 0.5$, the DFARIMA method using AIC has sample mean of \hat{d} close to zero and negative sample mean for \hat{c} , although the estimate of c is still positively biased. The sample mean of the estimate of c using BIC is still close to zero and large positive bias can be seen for the estimate of d . As α increases the use of BIC becomes preferred in terms of mean and standard deviation.

Polynomial $\alpha \leq 0.5$		Polynomial $\alpha > 0.5$	
Method	Mean \sqrt{MSE}	Method	Mean \sqrt{MSE}
FARIMA (AIC)	0.344	DFARIMA (BIC)	0.118
Local Whittle	0.346	Local Whittle	0.126
GPH	0.350	GPH	0.148
DFARIMA (AIC)	0.356	DFARIMA (AIC)	0.202
FARIMA (BIC)	0.385	FARIMA (BIC)	0.220
Dual Local Whittle	0.404	FARIMA (AIC)	0.374
DFARIMA (BIC)	0.408	Dual Local Whittle	0.471
DGPH	0.437	DGPH	0.482
Cusp		Logarithmic	
Method	Mean \sqrt{MSE}	Method	Mean \sqrt{MSE}
FARIMA (AIC)	0.238	DFARIMA (BIC)	0.257
GPH	0.275	Local Whittle	0.298
Local Whittle	0.305	GPH	0.309
FARIMA (BIC)	0.315	DFARIMA (AIC)	0.335
DFARIMA (AIC)	0.316	FARIMA (BIC)	0.351
DFARIMA (BIC)	0.391	FARIMA (AIC)	0.486
Dual Local Whittle	0.446	Dual Local Whittle	0.693
DGPH	0.453	DGPH	0.694

Tab. 5.15: Average \sqrt{MSE} for the various methods of estimating d in increasing order.

For the cusp maps the estimates of d for the DFARIMA methods are still very strongly positively biased. The sample means of the estimates of c are also generally positively biased, although negative bias is seen for the symmetric cusp map using AIC. The sample means of the estimates of c for the logarithmic maps are close to zero using both AIC and BIC and the sample means of the estimates of d are similar in magnitude as those for the standard estimation techniques shown in section 4.1. The sample means of the estimates of d for the logarithmic maps still show dependency on β , decreasing as β increases.

In comparison with the FARIMA methods, the sample standard deviations of

the DFARIMA methods using the same criterion are generally smaller. The sample standard deviations for the DFARIMA method using AIC are only larger than the FARIMA method using AIC for the polynomial map with $\alpha = 0.9$ and the logarithmic maps with $\beta = 0.25$ and 0.3 . The sample standard deviations for the DFARIMA method using BIC are only larger than the FARIMA method using AIC for the polynomial map with $\alpha = 0.9$ and equal correct to 2d.p. for the polynomial map with $\alpha = 0.8$. The sample standard deviations for the DFARIMA method using BIC for the logarithmic maps are the smallest of all the methods considered here and in section 4.1.

Table 5.15 gives the average \sqrt{MSE} for the new dual parameter estimation methods of estimating d in increasing order. The average \sqrt{MSE} for the standard local Whittle, GPII and FARIMA estimates are included for comparison. The dual local Whittle and DGPII method are seen to perform badly in terms of \sqrt{MSE} for all groups of maps. The standard deviations for these estimates are very large and the sample biases appear negatively biased. These methods thus seem unsuitable for the chaotic intermittency maps.

For the polynomial map with $\alpha \leq 0.5$ and the cusp maps the FARIMA methods outperform the new DFARIMA methods, although as mentioned in section 4.1, this is generally caused by the FARIMA methods failure to converge whilst the DFARIMA methods converge more on the wrong model. For order of performance is the same for the polynomial maps with $\alpha > 0.5$ and the logarithmic maps. In both cases the DFARIMA method using BIC outperforms the other estimation methods in terms of \sqrt{MSE} . Comparison with table 4.1 shows the DFARIMA method using BIC also performs better in terms of \sqrt{MSE} than the FAR and FExp methods for the logarithmic maps and only the FExp method gives better performance for the polynomial maps with $\alpha > 0.5$. The DFARIMA method using AIC also outperforms the FARIMA method using both AIC and BIC for these two groups of maps.

6. STOCHASTIC INTERMITTENCY MAPS

The estimates of d found in section 4.1 assume the map generating the orbits are unknown and attempt to find the rate of decay of the correlations using standard long memory parameter estimation techniques. If the generating map is known, however, the asymptotic decay rate of the autocorrelations is also known. Hence an alternative method of finding the rate of decay of the correlations for each orbit is to discover from which map the orbit was generated. For the deterministic chaotic intermittency maps described in section 3.2, this becomes a trivial problem. This section introduces new stochastic versions of the maps in which this problem is no longer trivial.

The study of stochastic versions of chaotic maps has been carried out previously by several authors. For example, Chan and Tong [1994], [2001], discuss how due to factors such as measurement error, even when a chaotic map provides a useful model for a physical phenomenon, the fit of a deterministic map would rarely be suitable. Lawrance and Hilliam [2005] consider methods of reducing noise in chaotic communications by means of a distribution transformation.

Even without the presence of error, if a chaotic system is purely deterministic but the exact form is unknown, it may not be conceivable that the specified class of models would contain the true model, and thus the modelling of chaotic data with noise is considered a more robust and practical approach.

In the study of stochastic versions of chaotic maps, it is important to ensure the random element does not alter the fundamental properties of the maps. Work done by Alves and Arujo [2000], Alves and Viana [2002] and Alves et al [2004] study the effects of noise on chaotic systems and the conditions required such that as the level of noise decreases the system approaches that of the original deterministic system without noise.

The new stochastic versions of the intermittency maps introduced here are shown to possess the same properties as the deterministic maps in the laminar region and hence retain the asymptotic rates of decay discussed in section 3.2. Section 6.1 introduces the stochastic polynomial map and section 6.2 introduces the stochastic logarithmic map. The generalised stochastic polynomial-logarithmic map is introduced in Section 6.3 and Section 6.4 then presents the results of a simulation study of these stochastic maps. The forms of the stochastic maps presented here were suggested by Dr M. Holland and have been included and studied in this chapter with his permission.

6.1 Stochastic Polynomial Map

For the polynomial map, the asymptotic value of d for an orbit is given by

$$d = 1 - \frac{1}{2\alpha}, \quad (6.1)$$

see Sarig [2002] and Gouëzel [2004b]. This implies that estimation of d is equivalent to estimating α . If the orbit is known to be generated from the polynomial map this becomes a trivial problem, with

$$\alpha = \frac{\log\left(\frac{w_{t+1}}{w_t} - 1\right)}{\log(2w_t)}$$

for all $w_t < 0.5$.

This problem becomes non-trivial with the introduction of a stochastic element to the map. Here, as suggested by Dr M. Holland, the stochastic polynomial map is defined over the range $[0,1]$ by

$$\zeta_\alpha(w) = \begin{cases} w(1 + e^{u-\theta}w^\alpha) & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases} \quad (6.2)$$

where θ is a positive constant and u is a realisation of a random variable, U , with zero mean. This replaces the multiplicative term 2^α in the deterministic polynomial map with the random element $e^{u-\theta}$.

The choice of making the map stochastic in this manner is not unique and was decided upon for several reasons. Firstly, consider the use of an additive noise term, A , such that

$$\tilde{\zeta}_\alpha^+(w) = \begin{cases} w(1 + (2w)^\alpha) + A & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

To ensure the orbit remained within the region of $[0,1]$, the distribution of A must be such that $-w \leq A \leq 1 - w$ and therefore such a term would need to be dependent on the value of w_t at each iteration, as independence would imply $A = 0$. Further, when the orbit entered the laminar region, this additive term could immediately return the orbit to the chaotic region on the next iteration and would therefore remove the intermittent nature of the deterministic map, altering the maps properties. Thus a multiplicative term is used. However, had this multiplicative term been placed outside the brackets, further difficulties may also occur. For example, consider the map

$$\tilde{\zeta}_\alpha^*(w) = \begin{cases} Aw(1 + (2w)^\alpha) & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

where A is a random variable. If $A < 0$ or $A > 1$ the map could escape the region of $J = [0, 1]$, hence the distribution of A must be such that $A \in [0, 1]$. This implies that multiplication by A will reduce the size of the mapped value. Note, of particular concern, is that once an orbit entered the laminar region it may remain there possibly indefinitely. This would greatly alter the properties of the map from those of the deterministic case.

Finally, it was decided for simplicity that the random variable would be independent of α . Note, that writing the random element in the form $e^{u-\theta}$ is equivalent to use of a random variable A , with $u = \log(A) + \theta$ and does not alter the map. This notation is used to simplify the equations when it comes to estimating α in section 6.1.1.

Now, for the stochastic polynomial map as defined by equation 6.2, the distribution of U must be such that the map remains within the region $[0, 1]$ for all $w \in [0, 1]$ and all $\alpha \in [0, 1]$, hence the distribution of U must be such that

$$\begin{aligned} w(1 + e^{u-\theta}w^\alpha) &\leq 1 \\ \Leftrightarrow u &\leq \theta + \log(1-w) - (1+\alpha)\log(w) \\ &= \theta + \alpha \log(2) \leq \theta \quad \text{for } \alpha \in [0, 1], \end{aligned} \quad (6.3)$$

that is, the upper bound of U is given by θ . Note, this upper bound implies $w_{t+1} = 1$ can only be achieved when $w_t = 0.5$ and $\alpha = 0$ and $u = \theta$, hence the general stochastic polynomial map will not cover the full range of $J = [0, 1]$ for $\alpha > 0$.

The following theorem gives the conditional mean and variance of w_1 given w_0 . It is proved for a more general case in section 6.3.

Theorem 6.1.1. *Let $E(e^{U-\theta}) = A$, where $0 < A < 1$ and $E(e^{2(U-\theta)}) = B$, where $A^2 < B < 1$. Given an initial value $w_0 < 0.5$, the value w_1 generated from w_0 using the stochastic polynomial map will have expected value, m_1 , generated by the deterministic map*

$$\zeta_{\alpha,A}(m_0) = m_0(1 + Am_0^\alpha) = m_1 \quad 0 \leq m_0 \leq 1/2 \quad (6.4)$$

and variance, v_1 , generated by the deterministic map

$$\zeta_{\alpha,A,B}(v_0) = v_0^{2(\alpha+1)}(B - A^2) = v_1 \quad 0 \leq v_0 \leq 1/2$$

with initial values $m_0 = v_0 = w_0$.

Due to the chaotic nature of the maps, a general extension of Theorem 6.1.1 to include the conditional means of $\{w_t\}$ given w_0 for $t \geq 2$ would become increasingly complicated as t increased and of little practical value. However, the following theorem uses Theorem 6.1.1 to give properties of an orbit generated from an initial value

arbitrarily close to zero such that there exists a small δ such that $w_t < \delta < 0.5$ for all $t \in (1, \dots, n)$. The proof can be found in section 6.3.

Theorem 6.1.2. *Let $0 \leq w_0 < 0.5$ be the initial value of an orbit w_1, \dots, w_n generated using the stochastic polynomial map such that $w_t < \delta < 0.5 \forall t \in (0, \dots, n)$. Then, as $w_0 \rightarrow 0$ and $\delta \rightarrow 0$, the orbit w_1, \dots, w_n will tend to the orbit m_1, \dots, m_n , generated by the deterministic map*

$$\zeta_{\alpha, A}(m) = m(1 + Am^\alpha) \quad 0 \leq m \leq 1/2$$

with initial value $m_0 = w_0$.

Theorem 6.1.2 shows that the behaviour of the stochastic polynomial map is the same as the deterministic polynomial map near the fixed point. Since the asymptotic rate of decay of the autocorrelations in the deterministic case are derived from this behaviour near the fixed point, this suggests that the asymptotic rate of decay of autocorrelations for the stochastic maps are the same as those for the deterministic polynomial maps.

6.1.1 Estimation of Alpha for the Stochastic Polynomial Map

Since the relationship given in equation 6.1 still appears to hold, the estimation of α for the stochastic polynomial map is equivalent to estimating the long memory parameter, d . Given an orbit w_1, \dots, w_n from the stochastic polynomial map,

$$w_{t+1} = w_t(1 + e^{u_t - \theta} w_t^\alpha), \quad \text{for } w_t \leq 0.5, \quad t \in (1, \dots, n-1).$$

Rearranging this and taking logs gives

$$\log\left(\frac{w_{t+1}}{w_t} - 1\right) = -\theta + \alpha \log(w_t) + u_t, \quad \text{for } w_t \leq 0.5, \quad t \in (1, \dots, n-1) \quad (6.5)$$

and standard linear regression techniques give the least squares estimate of α as

$$\hat{\alpha} = \frac{\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})^2}, \quad (6.6)$$

where $y_t = \log(w_{t+1}/w_t - 1)$, $x_t = \log(w_t)$ for $w_t \leq 0.5$ and $\tilde{n} \leq n-1$ is the number of observations, $w_t, t \in (1, \dots, n-1)$, such that $w_t \leq 0.5$.

Note, the series of $\{y_t\}$ and $\{x_t\}$ may not be consecutive values as they are obtained only for t such that $w_t \leq 0.5$. In the case when $w_t > 0.5$ they are disregarded as they possess no useful information of α . Once the orbit $\{w_t\}$ returns to below 0.5 the values of y_t and x_t are once again obtained. Hence only \tilde{n} values of y_t and x_t are available and the summations in equation 6.6 are over these available values.

Note also, that although the invariant density of the stochastic polynomial map is unknown, it is known that if $w_0 \neq 1$, the probability of the orbit returning to the state $w_t < 0$ tends to 1 as $t \rightarrow \infty$ and hence $\bar{n} \rightarrow \infty$ as $n \rightarrow \infty$.

The asymptotic distribution of $\hat{\alpha}$ is given in Theorem 6.1.3 under the following assumptions on U .

Assumption 6.1.1. $\{u_t\}$ for $t \in (1, \dots, n)$, are independent identically distributed random variables with $E(u_t) = 0$

Assumption 6.1.2. $\{u_t\}$ for $t \in (1, \dots, n)$, have distribution such that

$$\lim_{n \rightarrow \infty} \frac{(\sum_{t=1}^n E((a_t u_t)^3))^{1/3}}{(\sum_{t=1}^n E((a_t u_t)^2))^{1/2}} = 0$$

for any real finite constants $\{a_t\}$, $t \in (1, \dots, n)$.

Assumption 6.1.3. $\max(\{u_t\}) \leq \theta$ for $t \in (1, \dots, n)$

The first two assumptions are required to make use of Lyapunov's central limit theorem. The last is the upper limit of U as discussed in equation 6.3.

Theorem 6.1.3. *Let $\{w_t\}$ be an observed orbit of length n generated from the stochastic polynomial map with $\{u_t\}$ such that assumptions 6.1.1-6.1.3 hold, let $\bar{n} \leq n - 1$ be the number of observations, w_t , $t \in (1, \dots, n - 1)$, such that $w_t \leq 0.5$ and let $\hat{\alpha}$ be the least squares estimate of α given by equation 6.6. Then, as $\bar{n} \rightarrow \infty$, the conditional distribution of $\hat{\alpha}$ given $\{w_t\}$ is*

$$\left(\sum_{t=1}^{\bar{n}} (x_t - \bar{x})^2 \right)^{0.5} (\hat{\alpha} - \alpha) \rightarrow_d N(0, \sigma_U^2)$$

where σ_U^2 is the variance of the random variables u_t , $x_t = \log(w_t)$ for $w_t < 0.5$, $\bar{x} = \sum x_t / \bar{n}$ and all summations are taken over the range of available values of $\{x_t\}$ and $\{y_t\}$.

Proof. Substitution of $y_t = -\theta + \alpha x_t + u_t$ into equation 6.6 gives

$$\hat{\alpha} = \frac{\sum_{t=1}^{\bar{n}} (x_t - \bar{x})(-\theta + \alpha x_t + u_t - (-\theta + \alpha \bar{x} + \bar{u}))}{\sum_{t=1}^{\bar{n}} (x_t - \bar{x})^2},$$

which rearranges to give

$$\hat{\alpha} - \alpha = \frac{\sum_{t=1}^{\bar{n}} (x_t - \bar{x})(u_t - \bar{u})}{\sum_{t=1}^{\bar{n}} (x_t - \bar{x})^2} \quad (6.7)$$

$$= \sum_{t=1}^{\tilde{n}} a_t (u_t - \bar{u}),$$

where

$$a_t = \frac{(x_t - \bar{x})}{\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})^2} \quad \text{for } t \in (1, \dots, n).$$

Now, given observed $\{x_t\}$ for $t \in (1, \dots, n)$, $\{a_t\}$ are finite real constants. Therefore the random variables $z_t = a_t u_t$ have variance

$$\text{var}(z_t) = a_t^2 \sigma_U^2,$$

where σ_U^2 is the variance of each of the u_t . Thus under assumptions 6.1.1 and 6.1.2 the Lyapunov's central limit theorem states

$$\frac{\sum_{t=1}^{\tilde{n}} a_t (u_t - \bar{u})}{\left(\sum_{t=1}^{\tilde{n}} a_t^2 \sigma_U^2\right)^{1/2}} \rightarrow_d N(0, 1), \quad (6.8)$$

as $\tilde{n} \rightarrow \infty$. Substitution of equation 6.8 into equation 6.7 gives the required result. □

Remark 6.1.1. Since $(x_t - \bar{x})^2 \geq 0$ for all $t \in (1, \dots, n)$, with the $P((x_t - \bar{x})^2 = 0) \rightarrow 0$, this gives $E\left(\left(\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})^2\right)^{0.5}\right) \geq C\tilde{n}^{1/2}$ for some $C > 0$, as $\tilde{n} \rightarrow \infty$. Thus the estimated $\hat{\alpha}$ is $\tilde{n}^{1/2}$ consistent for α under assumptions 6.1.1-6.1.3.

6.2 Stochastic Logarithmic Map

For the logarithmic map, the asymptotic value of d is 0.5 for all $\beta \in (0, 2 \log(2) - 1)$. The knowledge that the orbit is generated from a logarithmic map therefore gives the asymptotic value of d without further knowledge of the parameter β required. However, the asymptotic value of the secondary long memory parameter, the newly introduced parameter c , is dependent on β and given by

$$c = \frac{1 + \beta}{2}, \quad (6.9)$$

see Bhansali and Holland [2008b] with reference to Zygmund [1988].

Introduction of a stochastic element to the map is made in a similar manner to that of the stochastic polynomial map. Here, the stochastic logarithmic map is defined over the range $[0, 1]$ by

$$\zeta_\alpha(w) = \begin{cases} w(1 + e^{u-\theta}w(-\log(w))^{1+\beta}) & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

where θ is a positive constant and u is a realisation of a random variable, U , with zero mean. Similar restrictions to the distribution of U must be made as with the stochastic polynomial map such that the map remains within the region $[0,1]$ for all $w \in [0, 1]$ and all $\beta \in (0, 2 \log(2) - 1)$, hence the distribution of U must be such that

$$\begin{aligned} w(1 + e^{u-\theta}w \log(1/w)^{1+\beta}) &\leq 1 \\ \Leftrightarrow u &\leq \theta + \log(1 - w) - 2 \log(w) - (1 + \beta) \log(\log(1/w)) \\ &\leq \theta + \log(2) - \log(\log(2)) \quad \text{for } \beta \in [0, 2 \log(2) - 1] \end{aligned} \tag{6.10}$$

Note, since $\log(2) - \log(\log(2)) \approx 1.06$, this is a weaker condition on the upper limit of U than that for the polynomial map given in equation 6.3. As with the stochastic polynomial map, a value of $w_{t+1} = 1$ is only achievable when $w_t = 0.5$, $\beta = 0$ and u takes its upper boundary value and hence the stochastic logarithmic map with $\beta > 0$ does not cover the full range of $J = [0, 1]$.

The following Theorems give the corresponding properties of the stochastic logarithmic map as those given for the stochastic polynomial map in section 6.1. The proofs can be found in the generalised case in section 6.3.

Theorem 6.2.1. *Let $E(e^{U-\theta}) = A$, where $0 < A < 1$ and $E(e^{2(U-\theta)}) = B$, where $A^2 < B < 1$. Given an initial value $w_0 < 0.5$, the value w_1 generated from w_0 using the stochastic logarithmic map will have expected value, m_1 , generated by the deterministic map*

$$\zeta_{\alpha,A}(m_0) = m_0(1 + Am_0(-\log(m_0))^{1+\beta}) = m_1 \quad 0 \leq m_0 \leq 1/2 \tag{6.11}$$

and variance, v_1 , generated by the deterministic map

$$\zeta_{\alpha,A,B}(v_0) = v_0^4(-\log(v_0))^{2(1+\beta)}(B - A^2) = v_1 \quad 0 \leq v_0 \leq 1/2$$

with initial values $m_0 = v_0 = w_0$.

Theorem 6.2.2. *Let w_0 be an initial value of an orbit w_1, \dots, w_n generated using the stochastic logarithmic map such that $w_t < \delta < 0.5 \forall t \in (0, \dots, n)$. Then as $w_0 \rightarrow 0$ and $\delta \rightarrow 0$, the orbit w_1, \dots, w_n will tend to the orbit m_1, \dots, m_n , generated by the deterministic map*

$$\zeta_{\alpha,A}(m) = m(1 + Am(-\log(m))^{1+\beta}) \quad 0 \leq m \leq 1/2$$

with initial value $m_0 = w_0$.

Similar to the stochastic polynomial map, Theorem 6.2.2 shows the asymptotic rate of decay of correlations for the stochastic logarithmic map appears to be the same as that for the deterministic logarithmic map and thus the asymptotic value of d equal to 0.5 and the relationship between β and c should still hold in the stochastic case.

6.2.1 Estimation of Beta for the Stochastic Logarithmic Map

The method of estimation of β for the stochastic logarithmic map is similar to that of the estimation of α for the stochastic polynomial map given in section 6.1.1. Given an orbit w_1, \dots, w_n from the stochastic logarithmic map,

$$w_{t+1} = w_t(1 + e^{u_t - \theta} w_t (-\log(w_t))^{1+\beta}), \quad \text{for } w_t \leq 0.5, \quad t \in (1, \dots, n-1).$$

Rearranging this and taking logs gives

$$\log\left(\frac{w_{t+1} - w_t}{w_t^2(-\log(w_t))}\right) = -\theta + \beta \log(-\log(w_t)) + u_t, \quad \text{for } w_t \leq 0.5, \quad t \in (1, \dots, n-1) \quad (6.12)$$

and standard linear regression techniques give the least squares estimate of β as

$$\hat{\beta} = \frac{\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{\tilde{n}} (x_t - \bar{x})^2}, \quad (6.13)$$

where $y_t = \log((w_{t+1} - w_t)/(w_t^2(-\log(w_t))))$, $x_t = \log(-\log(w_t))$ for $w_t \leq 0.5$ and $\tilde{n} \leq n-1$ is the number of observations, $w_t, t \in (1, \dots, n-1)$, such that $w_t \leq 0.5$. The asymptotic distribution of $\hat{\beta}$ is given in Theorem 6.2.3 under assumptions 6.1.1 and 6.1.2 and the following assumption on the upper limit of U .

Assumption 6.2.1. $\max(\{u_t\}) \leq \theta + \log(2) - \log(\log(2))$ for $t \in (1, \dots, n)$

The first two assumptions, as with the proof for estimating α from the stochastic polynomial map, are required to make use of Lyapunov's central limit theorem. The last is the upper limit of U as discussed in equation 6.10. The proof follows the same arguments as those for Theorem 6.1.3 and is thus omitted.

Theorem 6.2.3. *Let $\{w_t\}$ be an observed orbit of length n generated from the stochastic logarithmic map with $\{u_t\}$ such that assumptions 6.1.1, 6.1.2 and 6.2.1 hold, let $\tilde{n} \leq n-1$ be the number of observations, $w_t, t \in (1, \dots, n-1)$, such that $w_t \leq 0.5$ and let $\hat{\beta}$ be the least squares estimate of β given by equation 6.13. Then, as $\tilde{n} \rightarrow \infty$, the conditional distribution of $\hat{\beta}$ given $\{w_t\}$ is*

$$\left(\sum_{t=1}^{\bar{n}} (x_t - \bar{x})^2 \right)^{0.5} (\hat{\beta} - \beta) \rightarrow_d N(0, \sigma_U^2)$$

where σ_U^2 is the variance of the random variables u_t , $x_t = \log(-\log(w_t))$ for $w_t < 0.5$, $\bar{x} = \sum x_t / \bar{n}$ and all summations are taken over the range of available values of $\{x_t\}$ and $\{y_t\}$.

Similar to the estimate of α for the stochastic polynomial map, remark 6.1.1 still holds to show the estimate of β for the stochastic logarithmic map is $\sqrt{\bar{n}}$ -consistent for β .

6.3 Stochastic Polynomial-Logarithmic Map

The stochastic polynomial-logarithmic map generalizes the stochastic polynomial and stochastic logarithmic maps introduced in the sections 6.1 and 6.2. It is a stochastic version of the deterministic polynomial-logarithmic map defined in equation 3.20. Here, the stochastic polynomial-logarithmic map is defined over the range $[0,1]$ by

$$\zeta_\alpha(w) = \begin{cases} w(1 + e^{u-\theta} w^\alpha (-\log(w))^{1+\beta}) & \text{if } 0 \leq w \leq 1/2, \\ 2w - 1 & \text{if } 1/2 < w \leq 1. \end{cases}$$

where θ is a positive constant, $\alpha \in (0,1]$, $\beta \in [-1, 2\log(2) - 1]$, where for $\alpha = 1, \beta > 0$ and u is a realisation of a random variable, U , with zero mean. Similar restrictions to the distribution of U must be made as with the stochastic polynomial map such that the map remains within the region $[0,1]$ for all $w \in [0,1]$. From the arguments for the stochastic polynomial and logarithmic maps, it can be seen that if the map remains within this region when $w = 0.5$, $\alpha = 0$ and $\beta = -1$ it will remain for all other required values of w and β . Note, when $\beta = -1$, the map becomes a stochastic polynomial map and hence the upper limit of u is given by θ from equation 6.3. Also note, for the case when $\alpha = 1$, the map becomes a stochastic logarithmic map and β must be greater than zero to ensure the orbits are stationary.

In general, if β were allowed to take values less than -1, such that for $\beta^* < -1$, $\beta \in (\beta^*, 2\log(2) - 1)$; then the distribution of U must be restricted such that

$$\begin{aligned} 0.5(1 + e^{u-\theta} 0.5^0 \log(2)^{1+\beta^*}) &\leq 1 \\ \Leftrightarrow u &\leq \theta - (1 + \beta^*) \log(\log(2)). \end{aligned}$$

Since $\log(\log(2)) \approx -0.37$, this becomes a stronger condition on the upper limit of u than those for the stochastic polynomial or logarithmic maps and if $\beta^* < -1 + \theta / \log(\log(2))$ the upper limit of U becomes less than zero which contradicts the assumption that U is a zero mean random variable.

The following Theorems give the generalisations of the properties of the stochastic polynomial map given in section 6.1 and the stochastic logarithmic map given in section 6.2 for the stochastic polynomial-logarithmic map.

Theorem 6.3.1. *Let $E(e^{U-\theta}) = A$, where $0 < A < 1$ and $E(e^{2(U-\theta)}) = B$, where $A^2 < B < 1$. Given an initial value $w_0 < 0.5$, the value w_1 generated from w_0 using the stochastic polynomial-logarithmic map will have expected value, m_1 , generated by the deterministic map*

$$\zeta_{\alpha,A}(m_0) = m_0(1 + Am_0^\alpha(-\log(m_0))^{1+\beta}) = m_1 \quad 0 \leq m_0 \leq 1/2$$

and variance, v_1 , generated by the deterministic map

$$\zeta_{\alpha,A,B}(v_0) = v_0^{2(1+\alpha)}(-\log(v_0))^{2(1+\beta)}(B - A^2) = v_1 \quad 0 \leq v_0 \leq 1/2$$

with initial values $m_0 = v_0 = w_0$.

Proof. Let $m_0 = v_0 = w_0 < 0.5$. The expected value of w_1 given w_0 is obtained by

$$\begin{aligned} E(w_1|w_0) &= E(w_0(1 + e^{U-\theta}w_0^\alpha(-\log(w_0))^{1+\beta})|w_0) \\ &= w_0(1 + E(e^{U-\theta})w_0^\alpha(-\log(w_0))^{1+\beta}) \\ &= m_0(1 + Am_0^\alpha(-\log(m_0))^{1+\beta}) \\ &= m_1 \end{aligned}$$

as required. Similarly, the expected value of w_1^2 given w_0 is obtained by

$$\begin{aligned} E(w_1^2|w_0) &= E(w_0^2(1 + e^{U-\theta}w_0^\alpha(-\log(w_0))^{1+\beta})^2|w_0) \\ &= w_0^2 E(1 + 2e^{U-\theta}w_0^\alpha(-\log(w_0))^{1+\beta} + e^{2(U-\theta)}w_0^{2\alpha}(-\log(w_0))^{2(1+\beta)}|w_0) \\ &= w_0^2(1 + 2Aw_0^\alpha(-\log(w_0))^{1+\beta} + Bw_0^{2\alpha}(-\log(w_0))^{2(1+\beta)}), \end{aligned}$$

giving the variance of w_1 given w_0 as

$$\begin{aligned} Var(w_1|w_0) &= E(w_1^2|w_0) - E(w_1|w_0)^2 \\ &= w_0^2(1 - 1 + (2A - 2A)w_0^\alpha(-\log(w_0))^{1+\beta} + (B - A^2)w_0^{2\alpha}(-\log(w_0))^{2(1+\beta)}) \\ &= v_0^{2(1+\alpha)}(-\log(v_0))^{2(1+\beta)}(B - A^2) \\ &= v_1 \end{aligned}$$

as required. □

Remark 6.3.1. The constants A and B exist since $0 < e^{u-\theta} \leq 1$ and $0 < e^{2(u-\theta)} \leq 1$ because $u \leq \theta$.

Remark 6.3.2. The mean and variance of w_1 conditional on w_0 for $w_0 > 0.5$ is trivial, with $m_1 = 2m_0 - 1$ and $v_1 = 0$, since the stochastic element is only present when $w_0 \leq 0.5$.

Theorem 6.3.2. *Let w_0 be an initial value of an orbit w_1, \dots, w_n generated using the stochastic polynomial-logarithmic map such that $w_t < \delta < 0.5 \forall t \in (0, \dots, n)$. Then as $w_0 \rightarrow 0$ and $\delta \rightarrow 0$, the orbit w_1, \dots, w_n will tend to the orbit m_1, \dots, m_n , generated by the deterministic map*

$$\zeta_{\alpha, A}(m) = m(1 + Am(-\log(m))^{1+\beta}) \quad 0 \leq m \leq 1/2$$

with initial value $m_0 = w_0$.

Proof. Take $w_0 < \delta < 0.5$. Making use of theorem 6.3.1 gives the conditional expected value of w_1 as m_1 and the variance as $w_0^{2(1+\alpha)}(-\log(w_0))^{2(1+\beta)}(B - A^2) < \delta^{2(1+\alpha)}(-\log(\delta))^{2(1+\beta)}(B - A^2)$. Hence, as $\delta \rightarrow 0$, the variance of w_1 tends to zero and $w_1 \rightarrow m_1$.

Let $w_k \rightarrow m_k < \delta < 0.5$, with $k \in (1, \dots, n - 1)$, then, from theorem 6.3.1, the conditional expected value of w_{k+1} is given by m_{k+1} and the conditional variance is given by $m_k^{2(1+\alpha)}(-\log(m_k))^{2(1+\beta)}(B - A^2) < \delta^{2(1+\alpha)}(-\log(\delta))^{2(1+\beta)}(B - A^2)$. Hence, as $\delta \rightarrow 0$, the variance of w_{k+1} tends to zero and $w_{k+1} \rightarrow m_{k+1}$. Thus the proof is completed by induction. \square

6.3.1 Simultaneous Estimation of Alpha and Beta

As with the stochastic polynomial and logarithmic maps, estimates of α and β for the stochastic polynomial-logarithmic map can be found using linear regression techniques. Given an orbit w_1, \dots, w_n from the stochastic polynomial-logarithmic map,

$$w_{t+1} = w_t(1 + e^{u_t - \theta} w_t^\alpha (-\log(w_t))^{1+\beta}), \quad \text{for } w_t \leq 0.5, \quad t \in (1, \dots, n - 1).$$

Rearranging this and taking logs gives, for $w_t \leq 0.5$, $t \in (1, \dots, n - 1)$,

$$\log\left(\frac{w_{t+1} - w_t}{w_t(-\log(w_t))}\right) = -\theta + \alpha \log(w_t) + \beta \log(-\log(w_t)) + u_t, \quad (6.14)$$

and the least squares estimates of α and β by regression of $y_t = \log((w_{t+1} - w_t)/(w_t(-\log(w_t))))$ onto $x_{1t} = \log(w_t)$ and $x_{2t} = \log(-\log(w_t))$ for $w_t \leq 0.5$, that is

$$\hat{\Gamma} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}, \quad (6.15)$$

where $\hat{\Gamma} = (\hat{\theta}, \hat{\alpha}, \hat{\beta})'$, Y is the vector of y_t and X is the matrix with columns of 1's, x_{1t} and x_{2t} . As previously there are \tilde{n} values of $\{y_t\}$ and X is a $\tilde{n} \times 3$ matrix, where $\tilde{n} \leq n - 1$ be the number of observations, w_t , $t \in (1, \dots, n - 1)$, such that $w_t \leq 0.5$. Under assumptions 6.1.1-6.1.3, Lyapunov's central limit theorem and the arguments of theorems 6.1.3 and 6.2.3 suggests the conditional distribution of these estimates to be approximately of the form

$$(\hat{\Gamma} - \Gamma_0) \approx N(0, \sigma_U^2 (X'X)^{-1}) \quad (6.16)$$

for large finite \tilde{n} , where σ_U^2 is the variance of the random variables u_t .

A rigorous asymptotic proof of the distribution is not straightforward and left for future works. The difficulty arises due to the regression on two slowly varying functions of w_t , with the density of w_t itself unknown. The results of Phillips [2001] are not directly applicable in this case since the regressors are not smooth, but they suggest the matrix of regressors may be asymptotically singular.

Interestingly, recall that estimating α and β is equivalent to estimating d and c respectively and note the similarities between the regression given in equation 6.5 and the regression required for the DGPII method given in equation 5.6. The method presented here is a time domain approach specifically designed for the stochastic polynomial-logarithmic map whilst the DGPII method is a spectral approach applicable to a more general set of processes, but the difficulty of asymptotic multicollinearity may be common to both.

6.4 Empirical Studies

In order to find empirical evidence to support the theorems of section 6 for the stochastic polynomial, logarithmic and polynomial-logarithmic maps, several simulation studies are now carried out. First, empirical evidence is sought to support theorems 6.1.2, 6.2.2 and 6.3.2, that is, that the stochastic maps behaviour near the neutral fixed points tends to that of the deterministic maps. This is important, since the relationships between the map parameters α and β and the long memory parameters d and c depend on these properties and thus estimation of α and β can only be considered equivalent to estimating d and c if these theorems hold.

Having looked at the laminar region behaviour of the stochastic maps, studies are then carried out to estimate α and β from simulations of the maps, and to compare the distributions of these estimates with the theoretical ones given in theorems 6.1.3 and 6.2.3 and equation 6.16. As these results assume the type of map has been correctly specified, a final simulation study is then carried out to test if the correct map can be fitted when assumed unknown.

6.4.1 Laminar Region Behaviour

Theorems 6.1.2, 6.2.2 and 6.3.2 show that orbits from the stochastic versions of the polynomial, logarithmic and polynomial-logarithmic maps tend towards the corresponding orbits from the deterministic versions of the maps with the same parameters. This result is essential for the assumption that the parameters of the stochastic maps have the same relationship with the asymptotic long memory parameters of the orbits as those for the deterministic case, since it is the behaviour near the neutral fixed point which these relationships are based on for the deterministic case.

Since theorems 6.1.2, 6.2.2 and 6.3.2 are applicable as $w_0 \rightarrow 0$, seven values of w_0 were considered, namely $w_0 \in \{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, 10^{-11}\}$. The values of α and β included in the study were the same as those used in the study of the deterministic maps, that is $\alpha \in \{0.3, 0.45, 0.5, 0.65, 0.8, 0.9\}$ and $\beta \in \{0.05, 0.15, 0.25, 0.3\}$. The parameter θ was taken such that $\theta \in \{0.5, 1, 1.5\}$ and the distribution of U was taken as uniform, such that $U \sim \text{Uni}(-\theta, \theta)$ for a chosen θ .

This choice of U was made such that U is a zero mean random variable which satisfies assumptions 6.1.1-6.1.3. From Theorem 6.1.1, the conditional one step ahead variance of w_1 given w_0 is given by

$$\text{Var}(w_1|w_0) = w_0^{2(\alpha+1)} \text{Var}(e^{U-\theta}).$$

Now, for $U \sim \text{Uni}(-\theta, \theta)$,

$$\text{Var}(e^{U-\theta}) = \int_{-\theta}^{\theta} \frac{e^{2(U-\theta)}}{2\theta} dU - \left(\int_{-\theta}^{\theta} \frac{e^{U-\theta}}{2\theta} dU \right)^2,$$

which, by standard integration techniques, gives

$$\text{Var}(e^{U-\theta}) = \frac{\theta - 1 + 2e^{-2\theta} - (1 + \theta)e^{-4\theta}}{4\theta^2}.$$

Figure 6.1 gives a plot of $\text{Var}(e^{U-\theta})$ against θ , showing as $\theta \rightarrow 0$ the variance tends to zero, since the random variable U tends to a constant zero, but also as θ gets larger than 1.62 the variance begins to decrease again, since $e^{U-\theta} = e^U e^{-\theta}$ and $e^{-\theta} \rightarrow 0$ as $\theta \rightarrow \infty$. The theoretical effect of θ on the variance after multiple iterations becomes more complicated. Use of Theorem 6.2.1 and Theorem 6.3.1 show similar results would be seen for the stochastic logarithmic and stochastic polynomial-logarithmic maps. Hence, no obvious trends due to θ are expected from the results of these maps for this form of U .

For the stochastic polynomial map, 1000 orbits each of length $T = 1000$ were generated for each of the $7 \times 6 \times 3$ possible $\{w_0, \alpha, \theta\}$ combinations. An orbit of length $T = 1000$ generated from the deterministic polynomial map given in equation 6.4

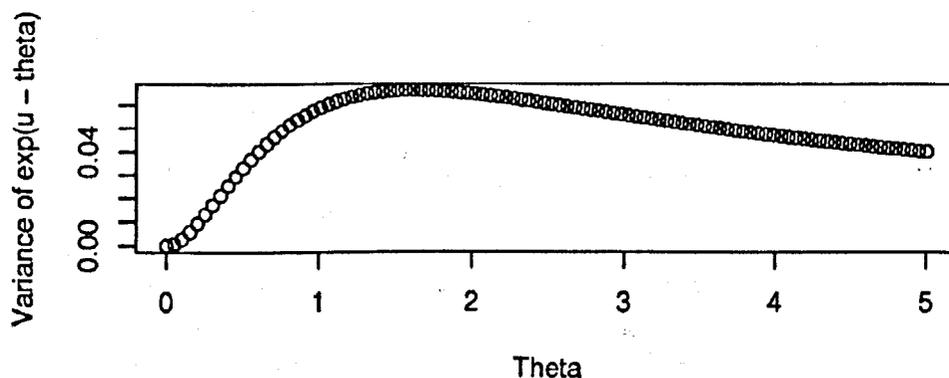


Fig. 6.1: The variance of $e^{U-\theta}$ against θ for $U \sim \text{Uni}(-\theta, \theta)$.

is also simulated for each of the $7 \times 6 \times 3$ possible $\{w_0, \alpha, \theta\}$ combinations. The stochastic orbits are referred to as $\{w_t\}$, whilst the corresponding deterministic orbits are referred to as $\{m_t\}$.

Similarly, for the stochastic logarithmic map, 1000 orbits each of length $T = 1000$ were generated for each of the $7 \times 4 \times 3$ possible $\{w_0, \beta, \theta\}$ combinations and a deterministic orbit of length $T = 1000$ generated from equation 6.11 corresponding to each choice of parameters was also simulated. For the stochastic and deterministic polynomial-logarithmic maps, two values of α and β were used, $\alpha \in (0.45, 0.8)$ and $\beta \in (0.05, 0.3)$, and 1000 stochastic orbits and 1 deterministic orbit each of length $T = 1000$ were generated for each of the $7 \times 2 \times 2 \times 3$ possible $\{w_0, \alpha, \beta, \theta\}$ combinations.

To compare the stochastic orbits with the corresponding deterministic orbits, the ratios w_t/m_t for $t \in (1, \dots, 1000)$ were obtained for each simulated orbit. The theoretical results suggest this ratio should tend to 1 as $w_0 \rightarrow 0$.

The results for the stochastic polynomial map with $w_0 = 10^{-5}$ and $\theta = 1$ are presented in figure 6.2. The thick center lines represent the mean of the ratios of w_t/m_t for each $t \in (1, \dots, 1000)$, whilst the thinner lines represent the mean plus and minus one standard deviation.

It can be seen that the standard deviations of these ratios decrease as α increases and that, for $\alpha \geq 0.5$, the mean ratio is close to 1 for all $t \in (1, \dots, 1000)$. Indeed, for $\alpha \geq 0.8$, the variability of the ratios is too small to be seen on this scale, with the maximum standard deviation for $\alpha = 0.8$ being 0.00077 and the maximum standard deviation for $\alpha = 0.9$ being 0.00026, but the scales have been kept constant for ease of comparison between the varying values of α . For $\alpha < 0.5$, many of the simulated orbits escaped from the laminar region for this choice of $w_0 = 10^{-5}$ and thus Theorem

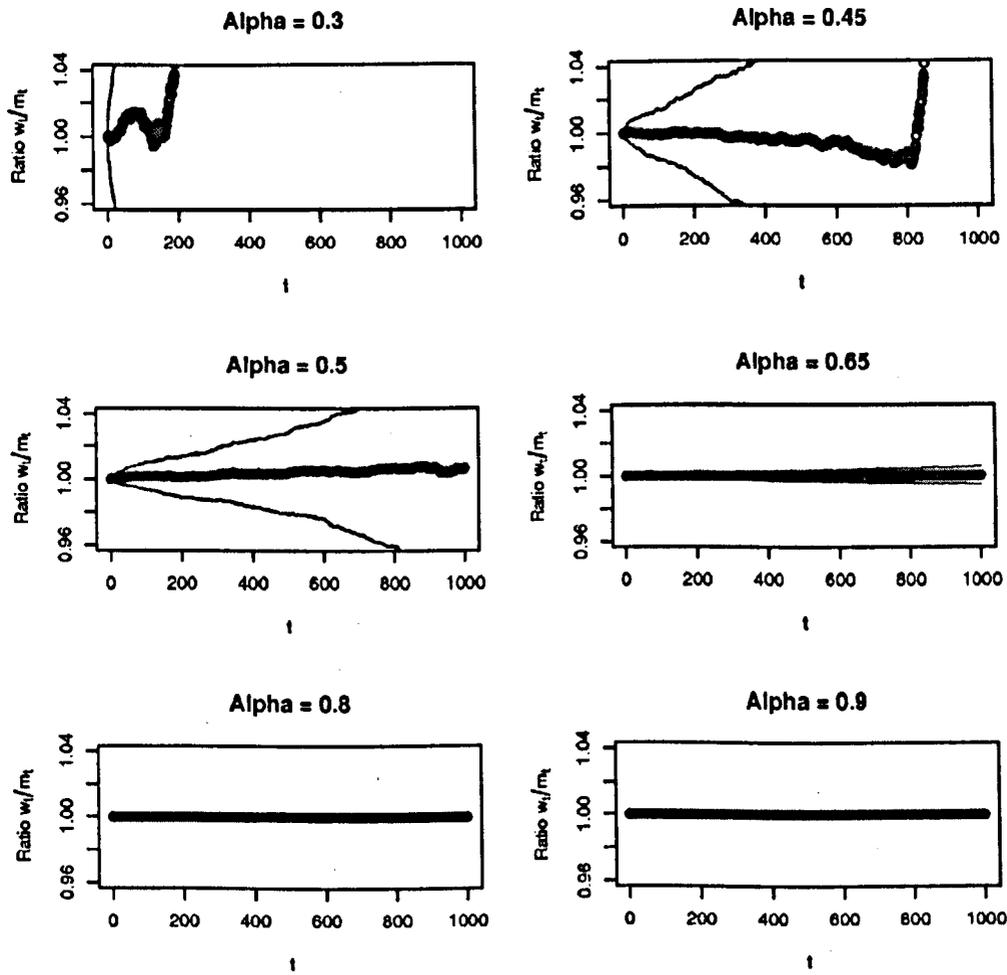


Fig. 6.2: The mean ratios of w_t/m_t , $t \in (1, \dots, 1000)$, for the stochastic polynomial map with $w_0 = 10^{-5}$ and $\theta = 1$. The thinner lines represent plus and minus one standard deviation.

6.1.2 does not apply to these orbits. For the initial part of the orbits which remain in the laminar region, however, the mean ratios are still close to 1 as expected.

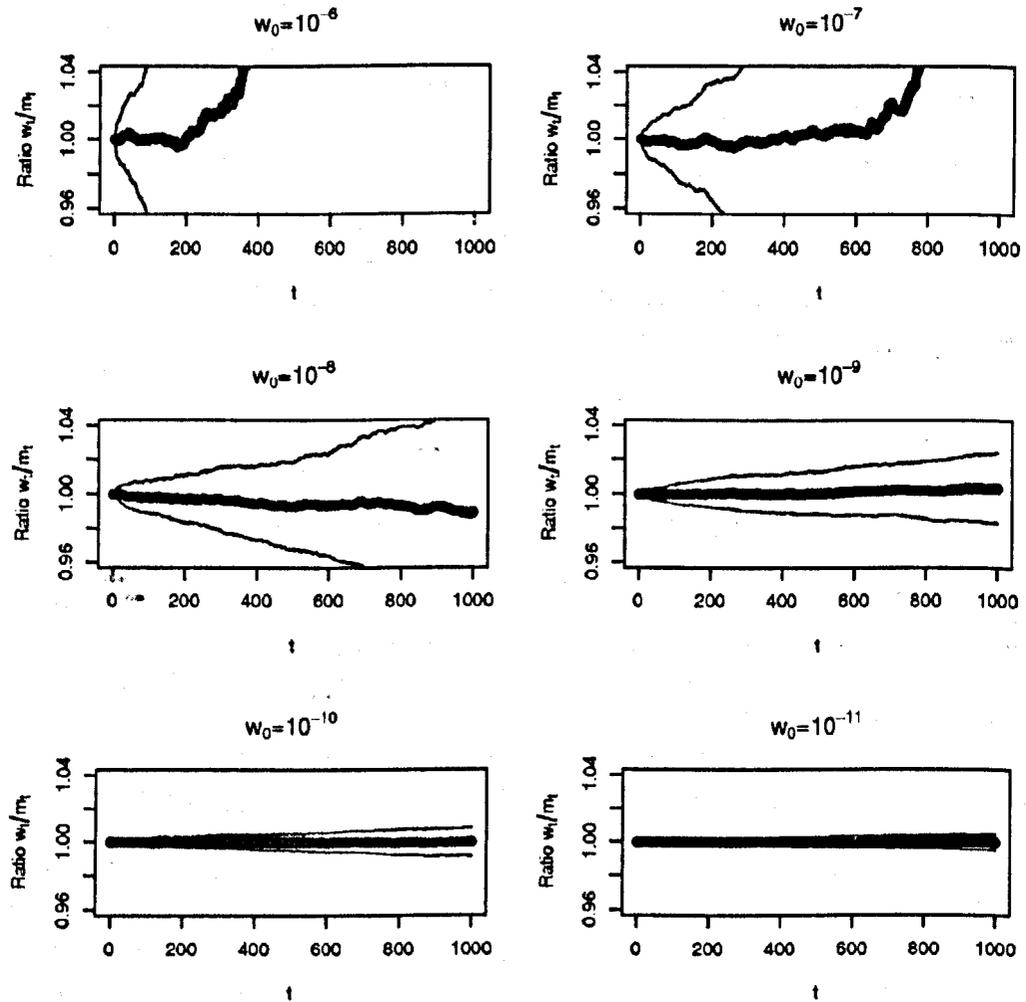


Fig. 6.3: The mean ratios of w_t/m_t , $t \in (1, \dots, 1000)$, for the stochastic polynomial map with $\alpha = 0.3$ and $\theta = 1$. The thinner lines represent plus and minus one standard deviation.

Figure 6.3 gives the results for the for the stochastic polynomial map with $\alpha = 0.3$ and $\theta = 1$ and $w_0 \in (10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, 10^{-11})$. The thick center lines once again represent the mean of the ratios of w_t/m_t for each $t \in (1, \dots, 1000)$, whilst the thinner lines represent the mean plus and minus one standard deviation. As the initial value w_0 decreases, the standard deviation of the ratios decrease and the mean ratios seem to tend to 1 as the theoretical results suggested.

As expected, the effect of θ does not appear consistent. Figure 6.4 shows the

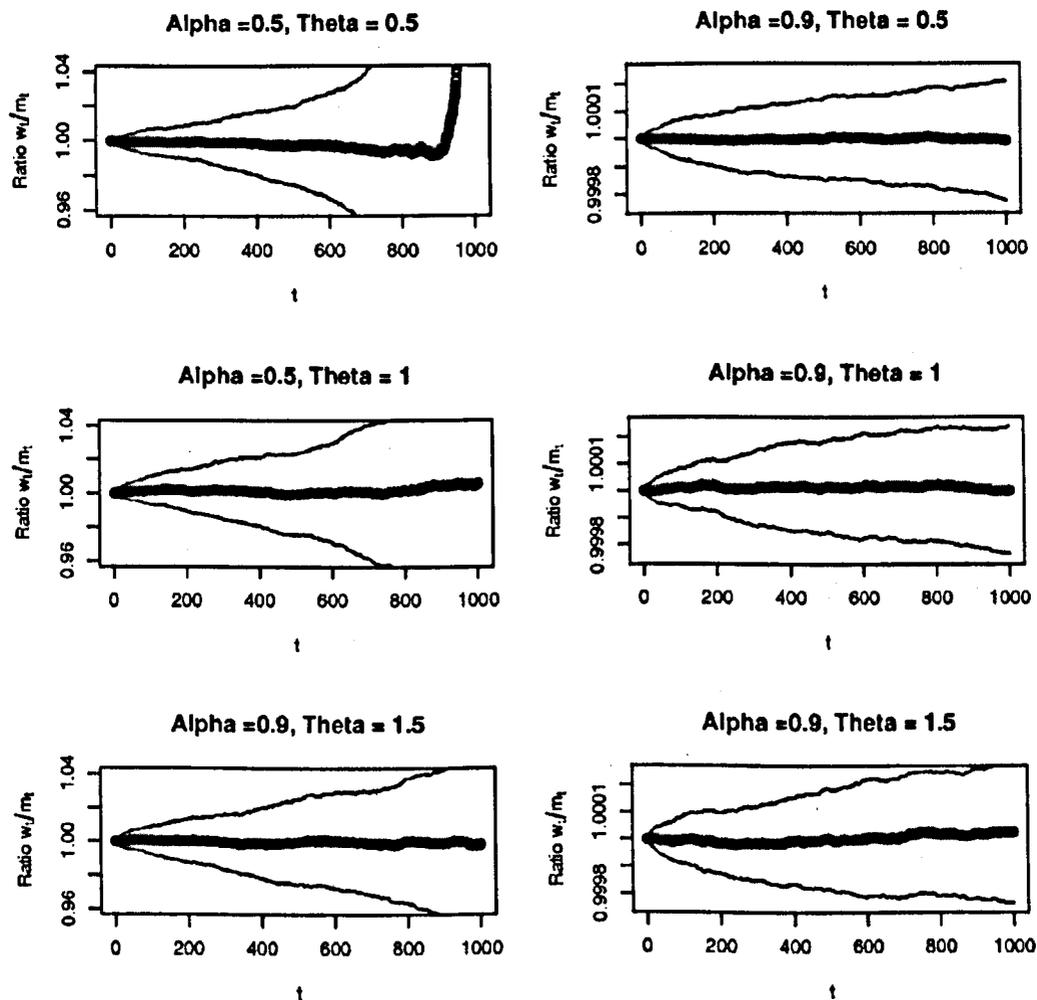


Fig. 6.4: The mean ratios of w_t/m_t , $t \in (1, \dots, 1000)$, for the stochastic polynomial map with $\alpha = 0.5$ and 0.9 and $w_0 = 10^{-5}$. The thinner lines represent plus and minus one standard deviation.

means of the ratios w_t/m_t for $t \in (1, \dots, 1000)$, for the stochastic polynomial map with $\alpha \in (0.5, 0.9)$, $w_0 = 10^{-5}$ and $\theta \in (0.5, 1, 1.5)$. The mean of the ratios appears close to 1, which is in agreement with the theoretical results. However, the standard deviations of the ratios for the stochastic polynomial map with $\alpha = 0.5$ appear to decrease slightly as θ increases, whilst the standard deviations appear to increase slightly with θ when $\alpha = 0.9$.

The trend of the standard deviations of the ratios tending to zero whilst the mean of the ratios tends to one as $w_0 \rightarrow 0$ is common for all values of alpha and for each initial value the standard deviations of the ratios decrease as alpha increases. The remaining graphs for the stochastic polynomial map are thus omitted.

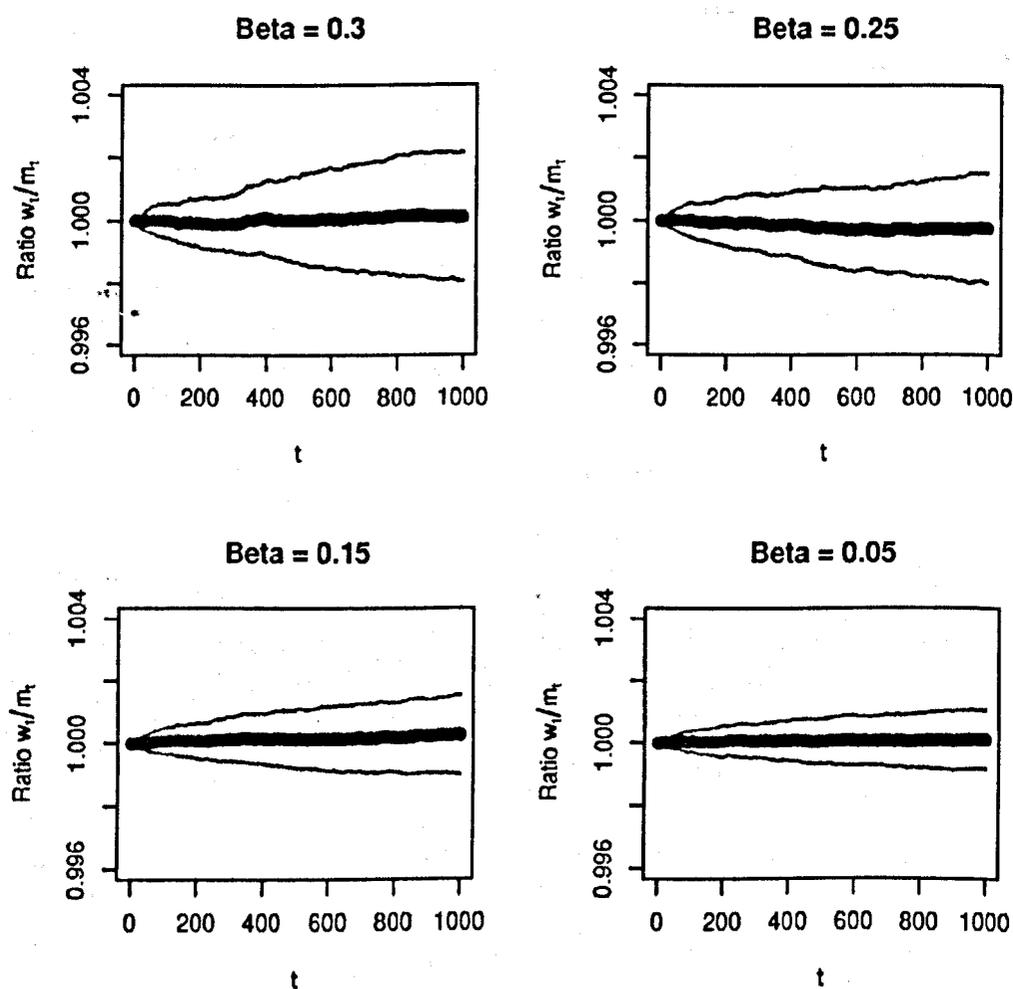


Fig. 6.5: The mean ratios of w_t/m_t , $t \in (1, \dots, 1000)$, for the stochastic logarithmic map with $w_0 = 10^{-5}$ and $\theta = 1$. The thinner lines represent plus and minus one standard deviation.

The results for the stochastic logarithmic map with $w_0 = 10^{-5}$ and $\theta = 1$ are presented in figure 6.5. As previously, the thick center lines represent the mean of the ratios of w_t/m_t for each $t \in (1, \dots, 1000)$, whilst the thinner lines represent the mean plus and minus one standard deviation.

For all values of β , the standard deviations of the ratios appear small and the means appear close to zero, as suggested by the theoretical results. Note the scale is smaller than the graphs for the stochastic polynomial maps. This follows from the pattern of the standard deviations decreasing as α increases for the stochastic polynomial map, as the stochastic logarithmic map is similar to a stochastic polynomial map with $\alpha = 1$.

The graphs show that as β decreases, the standard deviations of the ratios also decrease. This is true for all values of θ and all values of w_0 . In addition, as with the stochastic polynomial map, the standard deviations decrease further as $w_0 \rightarrow 0$.

For the generalised polynomial-logarithmic map, the results of both the stochastic polynomial and stochastic logarithmic maps can again be seen, in that for orbits which remain in the laminar region the means of the ratios w_t/m_t for $t \in (1, \dots, 1000)$ are close to 1 and the standard deviations of these ratios decrease as α increases, β decreases and $w_0 \rightarrow 0$. The effect of varying θ is once more not consistent. The graphs for these maps are similar in appearance to those already shown and thus omitted.

Overall, the empirical evidence from this study gives strong evidence to support theorems 6.1.2, 6.2.2 and 6.3.2, suggesting the stochastic maps behave as the corresponding deterministic maps near the neutral fixed point.

6.4.2 Estimation of Alpha and Beta

The previous study gave evidence to suggest the asymptotic relationships between the parameters α and β and the long memory parameters d and c established for the deterministic maps still hold for the new stochastic versions. These relationships are given in equations 6.1 and 6.9 and show that estimation of α and β is equivalent to estimating the asymptotic long memory parameters d and c .

Regression methods of estimating α and β from the stochastic polynomial, logarithmic and polynomial-logarithmic maps were presented in section 6, and the theoretical distributions of these estimates were presented in theorems 6.1.3 and 6.2.3 and equation 6.16. A simulation study is carried out here to test the fit of these theoretical distributions to finite sample estimates.

As in the previous simulation study, the values of α and β included in the study were the same as before, namely $\alpha \in \{0.3, 0.45, 0.5, 0.65, 0.8, 0.9\}$ for the stochastic polynomial map, $\beta \in \{0.05, 0.15, 0.25, 0.3\}$ for the stochastic logarithmic map and $(\alpha, \beta) \in ((0.45, 0.8), (0.05, 0.3))$ for the stochastic polynomial-logarithmic map. The parameter θ was once again taken such that $\theta \in \{0.5, 1, 1.5\}$ and the distribution of U was again uniform, such that $U \sim \text{Uni}(-\theta, \theta)$ for a chosen θ . These choices satisfy assumptions 6.1.1-6.1.3. The variance of this form of U given θ is $\sigma_U^2 = \theta^2/3$,

and from Theorems 6.1.3 and 6.2.3 and equation 6.16 this implies the variance of the estimates of α and β are expected to increase with θ .

The initial values for each simulated orbit, from all three stochastic maps, was generated from an exponential distribution truncated at 1, as with the deterministic case. A 'burn-in' time of $M = 10^7 - 10^4$ was then included, and the following $T = 10^4$ values were retained for study. The stochastic nature of the maps should place less importance on the choice of initial values and this 'burn-in' time may no longer be necessary. For each (α, θ) combination for the stochastic polynomial map, (β, θ) combination for the stochastic logarithmic map and (α, β, θ) combination for the stochastic polynomial-logarithmic map, 1000 orbits were simulated in this way.

Having retained an orbit of length $T = 10^4$, only the values of w_{t+1} and w_t such that $w_t < 0.5$ are used in the regression process. For comparison, only the first n such pair of values were taken from each orbit, where n was chosen as 100, 1000 and 4000. The upper value of 4000 for n was chosen as all of the simulated orbits had at least 4000 observations less than 0.5, whilst many did not have as many as 4500.

Estimated α for the Stochastic Polynomial Map							
α	θ	Bias			$\hat{\sigma}^2/\sigma^2$		
		n=100	n=1000	n=4000	n=100	n=1000	n=4000
0.3	0.5	0.002	-0.001	0.000	0.876	1.318	1.010
0.3	1	-0.005	-0.001	0.000	1.290	1.149	1.094
0.3	1.5	-0.011	0.002	-0.001	0.814	1.047	1.100
0.45	0.5	0.001	-0.001	0.000	0.955	0.824	1.047
0.45	1	0.005	-0.001	0.000	1.141	1.322	0.901
0.45	1.5	-0.001	-0.002	-0.001	1.049	0.943	1.028
0.5	0.5	0.004	0.000	0.000	1.092	0.908	0.661
0.5	1	-0.005	-0.003	-0.001	1.055	0.951	0.994
0.5	1.5	-0.003	-0.002	0.000	1.004	0.856	1.061
0.65	0.5	0.000	0.000	0.000	1.107	1.022	0.929
0.65	1	0.004	0.000	0.000	1.237	1.257	0.899
0.65	1.5	0.008	-0.003	-0.001	0.904	0.925	0.998
0.8	0.5	-0.002	-0.001	0.000	1.051	1.171	1.355
0.8	1	0.013	0.001	0.000	1.069	0.964	0.897
0.8	1.5	-0.003	-0.002	0.000	1.059	0.968	1.188
0.9	0.5	0.001	0.000	0.000	1.098	1.073	1.275
0.9	1	-0.014	-0.002	0.000	1.100	0.686	1.071
0.9	1.5	-0.023	0.000	0.000	1.188	0.970	1.130

Tab. 6.1: The mean biases and sample variances divided by the theoretical variances for the estimates of α from the stochastic polynomial map.

For the stochastic polynomial map an estimate of α was found for each value of n using equation 6.6, for the stochastic logarithmic map estimates of β were found

using equation 6.13 and for the stochastic polynomial-logarithmic map estimates of α and β were found simultaneously using equation 6.15.

Since the theoretical standard deviations are conditional on the value of $s_n = \left(\sum_{t=1}^n (\log(w_t) - \overline{\log(w_t)})^2\right)^{0.5}$, an estimate of this value was taken for each (α, θ, n) combination as

$$\bar{s}_n = \frac{1}{1000} \sum_{i=1}^{1000} s_{in},$$

where s_{in} was the value for the i th simulated orbit from a given (α, θ, n) combination. The theoretical standard deviations are then estimated using these values.

Table 6.1 gives the mean biases and sample variances divided by the theoretical variances given by Theorem 6.1.3 for the estimates of α from the stochastic polynomial map. It can be seen that even using only 100 observations, the biases are generally small and the sample variances are close to the theoretical ones in magnitude. When 4000 observations are used in the regression, biases are approximately zero to three decimal places. The results give strong empirical evidence in support of Theorem 6.1.3.

Figure 6.6 shows histograms of the estimates of α for the stochastic polynomial map with $\alpha = 0.3$. The theoretical distributions from Theorem 6.1.3 have been added on and the range of the x-axes for each histogram have been kept fixed for ease in comparison. It can be seen that the sample distributions appear close to the theoretical ones for all values of θ and n . It can also be seen that, as expected by the theoretical results, the sample variance increases as θ increases and decreases as n increases.

Theorem 6.1.3 shows no direct dependence of the distribution of $\hat{\alpha} - \alpha$ on the value of α . The sample distributions of the estimates of α agree with this and the results given in table 6.1 show no significant trend in terms of bias or variance of the estimates as α increases. The histograms for the remaining values of α are similar in appearance to those for $\alpha = 0.3$ and thus omitted.

Table 6.2 gives the mean biases and sample variances divided by the theoretical variances found from Theorem 6.2.3 for the estimates of β from the stochastic logarithmic map. These theoretical variances are estimated from the orbits as with the stochastic polynomial map. Once again the biases appear small even when $n = 100$ and the ratios of the sample variances with the theoretical values are close to 1 for most (β, θ, n) combinations. The case when $\beta = 0.15$ with $\theta = 1.5$ and $n = 100$ has an unusually high ratio of $\hat{\sigma}^2/\sigma^2 = 2.304$. This is caused by six large negative outliers of $\hat{\beta} \approx -2$, the reason for which appears to be all of the observed values used being less than 0.003 for these six orbits. Several of the other sample distributions for estimates of β and α also have one or two outliers for similar reasons when $n = 100$. Removal of these outliers for this case gives a ratio of $\hat{\sigma}^2/\sigma^2 = 1.044$.

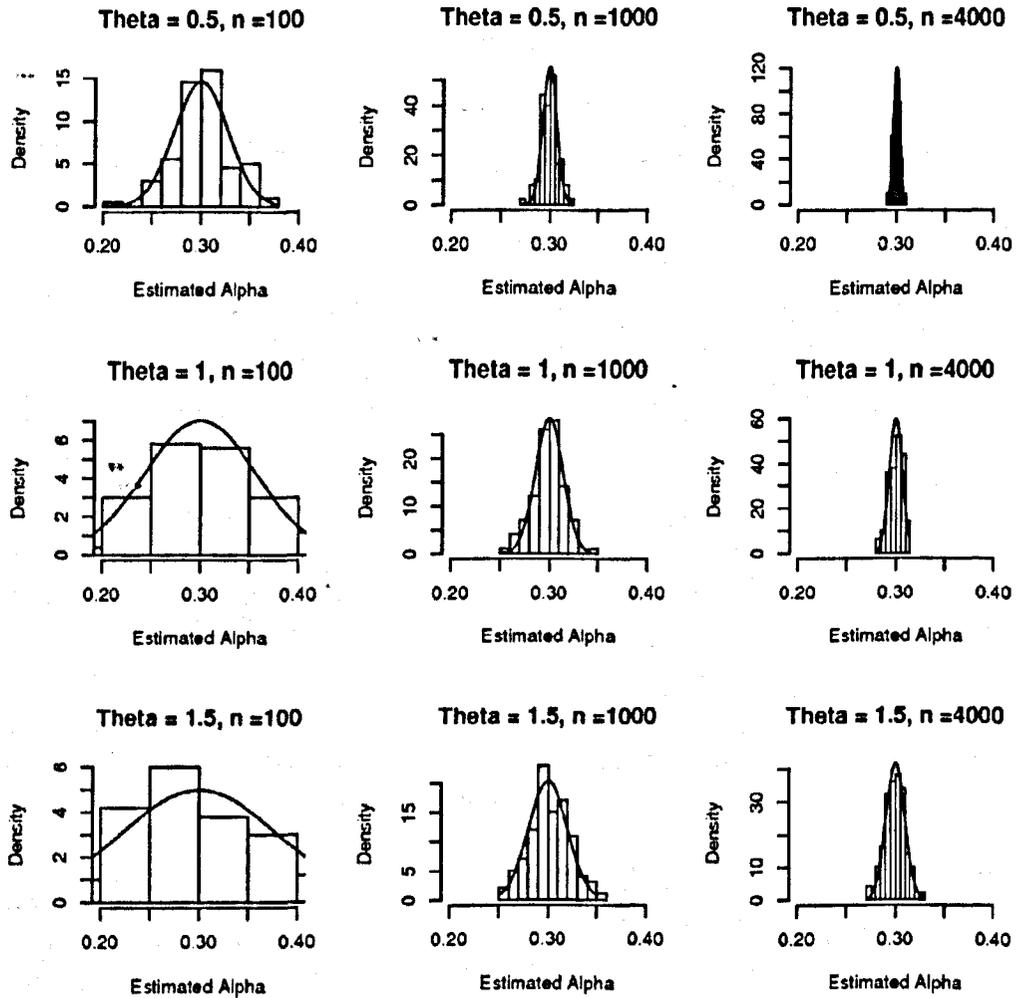


Fig. 6.6: Histograms of the estimates of α for the stochastic polynomial map with $\alpha = 0.3$. The lines represent the theoretical distributions.

Estimated β for the Stochastic Logarithmic Map							
β	θ	Bias			$\hat{\sigma}/\sigma$		
		n=100	n=1000	n=4000	n=100	n=1000	n=4000
0.05	0.5	-0.001	-0.001	0.001	1.530	0.854	1.147
0.05	1	0.013	-0.006	-0.003	1.286	1.132	0.922
0.05	1.5	-0.042	-0.002	0.002	1.371	0.988	1.197
0.15	0.5	-0.007	0.000	0.000	0.980	1.038	0.828
0.15	1	-0.012	-0.002	0.000	1.103	1.272	1.080
0.15	1.5	0.005	0.009	0.003	2.304	0.895	0.843
0.25	0.5	-0.008	-0.001	-0.001	1.203	0.904	1.063
0.25	1	-0.012	-0.003	-0.003	0.960	0.929	0.968
0.25	1.5	0.000	0.004	0.002	0.783	0.850	1.117
0.3	0.5	0.000	0.000	0.000	0.922	1.179	1.145
0.3	1	-0.001	-0.001	-0.001	1.206	0.830	1.075
0.3	1.5	0.000	0.007	-0.001	0.814	0.815	0.978

Tab. 6.2: The mean biases and sample variances divided by the theoretical variances for the estimates of β from the stochastic logarithmic map.

As with the stochastic polynomial map, the sample variances for the estimates of β from the stochastic logarithmic map decrease as n increases and increase as θ increases. The histograms for any given value of β are similar in appearance to those for the stochastic polynomial map shown in figure 6.6. As expected from Theorem 6.2.3, the mean biases and sample variances show no significant dependence on the value of β .

The mean biases and sample variances divided by the theoretical variances based on equation 6.16 for the estimates of α and β from the stochastic polynomial-logarithmic map are shown in table 6.3. As with the stochastic polynomial and logarithmic maps, the estimates of α and β from the generalised stochastic polynomial-logarithmic map show generally small biases and ratios of $\hat{\sigma}^2/\sigma^2$ close to 1. However, from table 6.1 it can be seen that for the estimates of α from the stochastic polynomial map, none of the ratios of $\hat{\sigma}^2/\sigma^2$ were larger than 1.4 and from table 6.2 it can be seen that for the estimates of β from the stochastic polynomial map, only two of the ratios of $\hat{\sigma}^2/\sigma^2$ were larger than 1.4 for $n = 100$ and none of the ratios were larger than 1.4 for $n \geq 1000$, whilst from table 6.3, for the estimates of α from the stochastic polynomial-logarithmic map three of the ratios of $\hat{\sigma}^2/\sigma^2$ were larger than 1.4 for $n = 100$, one of the ratios was larger than 1.4 for $n = 1000$ and one of the ratios was larger than 1.4 for $n = 4000$. The probability of outliers thus appears larger when α and β are estimated simultaneously.

Note also from table 6.3 that the signs of the biases of the estimates of α and β are generally the same and the magnitudes of the sample variances appear connected. This is due to strong correlation between the estimates of α and β from the

Estimated α for the Stochastic Polynomial-Logarithmic Map								
α	β	θ	Bias			$\hat{\sigma}/\sigma$		
			n=100	n=1000	n=4000	n=100	n=1000	n=4000
0.45	0.05	0.5	-0.003	-0.003	-0.002	1.092	1.132	0.839
0.45	0.05	1	0.020	-0.007	0.000	1.065	1.528	1.232
0.45	0.05	1.5	-0.063	0.001	-0.002	1.496	0.978	1.124
0.45	0.3	0.5	0.008	-0.002	-0.001	1.510	0.953	0.899
0.45	0.3	1	-0.062	-0.005	-0.001	0.962	0.990	1.077
0.45	0.3	1.5	0.012	-0.003	-0.006	0.635	1.223	0.996
0.8	0.05	0.5	0.001	-0.001	-0.002	1.254	0.885	1.124
0.8	0.05	1	-0.079	-0.004	0.000	1.111	1.311	0.943
0.8	0.05	1.5	-0.096	0.001	0.004	1.115	1.407	1.459
0.8	0.3	0.5	0.000	-0.005	0.000	0.884	1.239	0.850
0.8	0.3	1	0.064	-0.001	0.002	1.924	1.306	0.958
0.8	0.3	1.5	0.214	-0.005	-0.002	2.173	1.098	0.992
Estimated β for the Stochastic Polynomial-Logarithmic Map								
α	β	θ	Bias			$\hat{\sigma}/\sigma$		
			n=100	n=1000	n=4000	n=100	n=1000	n=4000
0.45	0.05	0.5	-0.007	-0.006	-0.005	0.951	1.082	0.937
0.45	0.05	1	0.068	-0.015	-0.001	0.893	1.378	1.206
0.45	0.05	1.5	-0.073	-0.002	-0.002	1.228	0.856	1.049
0.45	0.3	0.5	0.016	-0.004	-0.001	1.450	0.916	0.755
0.45	0.3	1	-0.133	-0.010	-0.004	0.889	0.852	1.111
0.45	0.3	1.5	0.034	-0.008	-0.013	0.672	1.115	0.994
0.8	0.05	0.5	-0.006	-0.003	-0.004	1.145	0.824	1.051
0.8	0.05	1	-0.156	-0.005	0.003	1.119	1.225	0.901
0.8	0.05	1.5	-0.198	0.012	0.012	0.939	1.369	1.409
0.8	0.3	0.5	-0.007	-0.010	0.000	0.885	1.243	0.863
0.8	0.3	1	0.243	-0.006	0.006	1.888	1.300	0.976
0.8	0.3	1.5	0.230	0.000	-0.002	2.158	1.008	0.970

Tab. 6.3: The mean biases and sample variances divided by the theoretical variances for the estimates of α and β from the stochastic polynomial-logarithmic map.

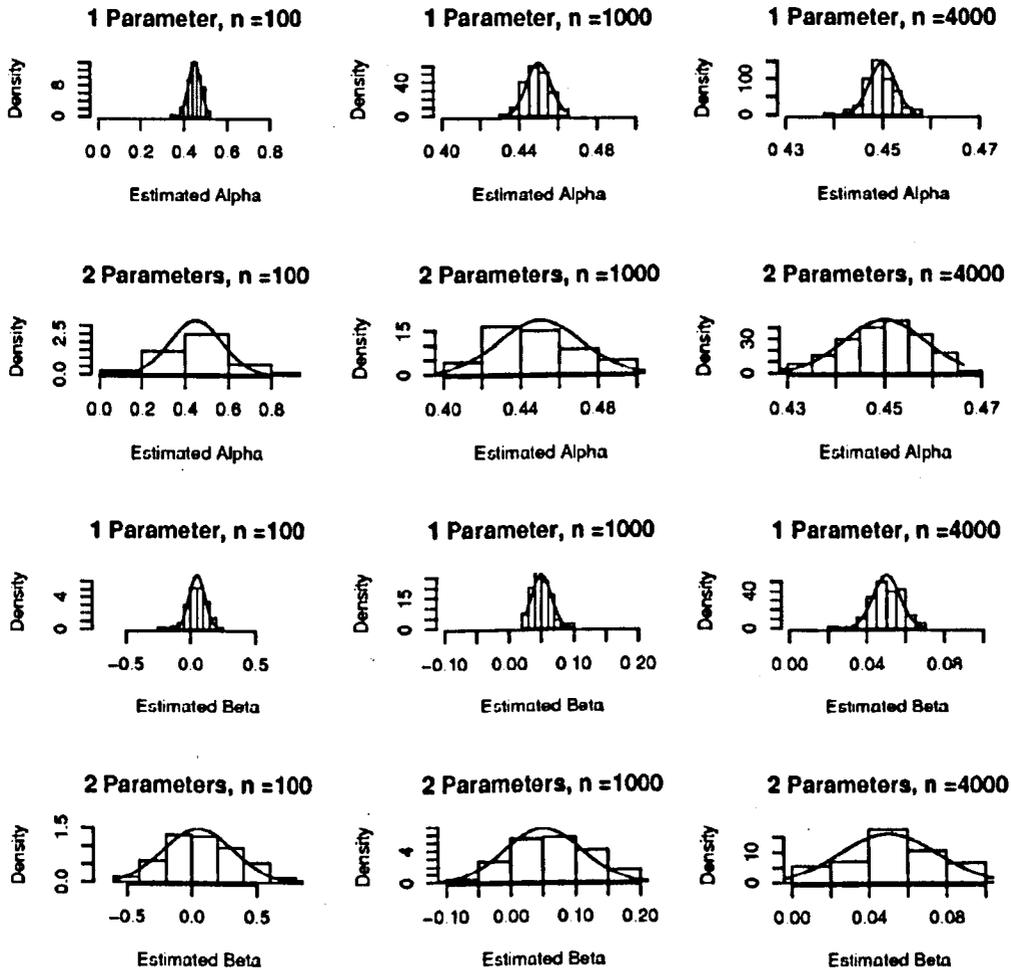


Fig. 6.7: Row 1: Histograms of the estimates of α from the stochastic polynomial map with $\alpha = 0.45$. Row 2: Histograms of the estimates of α from the stochastic polynomial-logarithmic map with $\alpha = 0.45$ and $\beta = 0.05$. Row 3: Histograms of the estimates of β from the stochastic logarithmic map with $\alpha = 0.05$. Row 4: Histograms of the estimates of β from the stochastic polynomial-logarithmic map with $\alpha = 0.45$ and $\beta = 0.05$. The lines represent the theoretical distributions and $\theta = 0.5$ for all.

stochastic polynomial-logarithmic map. Due to this correlation, the sample variances of the estimates of α and β estimated simultaneously are much larger than those of the estimates of α and β estimated separately from the stochastic polynomial and logarithmic maps. Examples of this increase in variance are shown in figure 6.7. Note that the results suggested by equation 6.16 still account for this increased variance and the theoretical distributions still fit the sample distributions well. As with the estimates of α and β from the stochastic polynomial and logarithmic maps, the values of α and β do not seem to alter the sample distributions from the stochastic polynomial-logarithmic map.

The results from this simulation study strongly support the theoretical results of theorems 6.1.3 and 6.2.3 and equation 6.16, suggesting that consistent estimates of α and β can be found for these stochastic maps, and hence consistent estimates of the asymptotic values of the long memory parameters d and c .

6.4.3 Discrimination between the Maps

The results of the simulation study presented in section 6.4.2 support theorems 6.1.3 and 6.2.3 and equation 6.16 and suggest consistent estimates of α and β can be found for these stochastic maps. However, these results assume it is known which of the stochastic maps an orbit is generated from. In this section this assumption is relaxed and two methods of model selection are introduced to determine if the correct generating map can be found from the data.

Figure 6.7 shows that although the stochastic polynomial-logarithmic map includes the stochastic polynomial and logarithmic maps as special cases, estimation of α and β simultaneously can greatly increase the variance of these estimates and should thus be avoided. Mistakenly fitting a stochastic polynomial map to a logarithmic map, or vice versa, will lead to bias, as the stochastic polynomial map has a value of $\beta = -1$, whilst the stochastic logarithmic map has a value of $\alpha = 1$. It is therefore important to correctly identify which of the maps an orbit is generated from.

The first approach to model selection is the use of a criterion such as AIC or BIC, see section 1.6.4. For a given orbit, all three regressions are carried out from equations 6.5, 6.12 and 6.14 and the sample standard deviations of the residuals are recorded for each. The criterion is then of the form

$$C = \log(\hat{\sigma}^2) + k \frac{p}{n},$$

where $\hat{\sigma}^2$ is the sample variance of the residuals, n is the number of observations used in the regression, p is the number of parameters such that $p = 1$ for the stochastic polynomial and logarithmic maps and $p = 2$ for the stochastic polynomial-logarithmic map and k is the penalty. The model chosen is then the one with the minimum value of C . For AIC $k = 2$ whilst for BIC $k = \log(n)$.

The second approach to model selection is to fit a stochastic polynomial-logarithmic map, estimating α and β simultaneously, and to then use the theoretical distribu-

tions of $\hat{\alpha}$ and $\hat{\beta}$ given by equation 6.16 to test the null hypotheses $H_{01} : \alpha = 1$ vs. $H_{11} : \alpha < 1$ and $H_{02} : \beta = -1$ vs. $H_{12} : \beta > 0$. If there is significant evidence to suggest $\alpha < 1$ and no significant evidence to reject $\beta = -1$ this suggests the orbit is generated from a stochastic polynomial map, whilst the reverse of these results would suggest a stochastic logarithmic map. If there is significant evidence to reject both null hypotheses this suggests the orbit was generated from a stochastic polynomial-logarithmic map. If there is no significant evidence to reject $\alpha = 1$ or $\beta = -1$ it suggests a special case of the stochastic-polynomial map. In terms of long memory, this would be equivalent to finding no significant evidence to reject $d = 0.5$ and $c = 0$ and thus that the orbit is non-stationary.

Such a finding would clearly be more likely with orbits generated from maps in which α was close to 1 and/or β was close to zero. Indeed, since the stochastic polynomial map has a value of $\beta = -1$, whilst the stochastic logarithmic map has a value of $\alpha = 1$, discrimination between the maps is expected to be less successful when α is close to 1 and/or β is close to zero whichever method is used. Note, however, that although it becomes more difficult to discriminate between the maps, the negative effect of fitting the wrong map is also reduced, since the bias of the estimate of α when a stochastic logarithmic map is fitted is $1 - \alpha$, whilst the bias of the estimate of β when a stochastic polynomial map is fitted is $\beta + 1$.

The simulation study carried out here follows the previous two. The values of α and β included in the study were the same as before, namely $\alpha \in \{0.3, 0.45, 0.5, 0.65, 0.8, 0.9\}$ for the stochastic polynomial map, $\beta \in \{0.05, 0.15, 0.25, 0.3\}$ for the stochastic logarithmic map and $(\alpha, \beta) \in ((0.45, 0.8), (0.05, 0.3))$ for the stochastic polynomial-logarithmic map. The parameter θ was once again taken such that $\theta \in \{0.5, 1, 1.5\}$ and the distribution of U was again $U \sim Uni(-\theta, \theta)$ for a chosen θ .

The initial values for each simulated orbit, from all three stochastic maps, was generated once more from an exponential distribution truncated at 1. A 'burn-in' time of $M = 10^7 - 10^4$ was then included, and the following $T = 10^4$ values were retained for study. For each (α, θ) combination for the stochastic polynomial map, (β, θ) combination for the stochastic logarithmic map and (α, β, θ) combination for the stochastic polynomial-logarithmic map, 100 orbits were simulated in this way.

For estimating α , β and σ^2 only the values of w_{t+1} and w_t such that $w_t < 0.5$ are used in the regression process. As before the first n such pair of values were taken from each orbit, where n was chosen as 100, 1000 and 4000.

For each simulated orbit, knowledge of the generating map was assumed unknown and the use of AIC, BIC and hypothesis testing was used to estimate which generating map best fit the orbit, as described above, with the hypothesis tests carried out at the 5% significance level. Table 6.4 gives the percentages that each of the selection methods chose each map for each group of orbits with $n = 100$ and $\theta = 1.5$. The effect of θ is discussed later, see table 6.7.

The results are less than satisfactory for this value of n . The correct model is

$n = 100, \theta = 1.5$										
	AIC			BIC			Hypothesis test			
Polynomial Map	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.3$	81	14	5	85	14	1	55	2	3	40
$\alpha = 0.45$	74	24	2	76	24	0	27	1	1	71
$\alpha = 0.5$	73	23	4	77	23	0	32	0	2	66
$\alpha = 0.65$	57	36	7	61	36	3	9	3	6	82
$\alpha = 0.8$	54	42	4	56	42	2	3	0	3	94
$\alpha = 0.9$	50	39	11	54	41	5	1	2	6	91
Logarithmic Map	α	β	Both	α	β	Both	α	β	Both	None
$\beta = 0.05$	32	67	1	32	67	1	3	15	1	81
$\beta = 0.15$	25	70	5	26	72	2	0	31	3	66
$\beta = 0.25$	27	69	4	27	73	0	6	32	0	62
$\beta = 0.3$	24	73	3	24	75	1	2	35	3	60
Poly-Log Map	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.45, \beta = 0.05$	52	34	14	60	36	4	38	19	13	30
$\alpha = 0.45, \beta = 0.3$	36	43	21	44	47	9	30	36	13	21
$\alpha = 0.8, \beta = 0.05$	42	56	2	43	57	0	12	27	1	60
$\alpha = 0.8, \beta = 0.3$	27	68	5	27	73	0	9	39	1	51

Tab. 6.4: the percentages the AIC, BIC and hypothesis testing methods chose each map for each group of orbits with $n = 100$ and $\theta = 1.5$. The headings α , β , 'Both' and 'None' refer to the method choosing the stochastic polynomial map, the stochastic logarithmic map, the stochastic polynomial-logarithmic map and the non-stationary stochastic polynomial-logarithmic map respectively.

fitted at best 85% of the time, though generally the success rate is less than 70 %. Correct identification of the stochastic polynomial-logarithmic map occurs at most 21% of the time, which occurs when $\alpha = 0.45$ is small, $\beta = 0.3$ is large and the model is selected by AIC. Note, even fitting one of the three models randomly would achieve a greater success rate of around 33%. The small sample effect thus makes discriminating between the maps difficult.

The use of AIC or BIC generally outperforms the use of the hypothesis tests for this orbit length, as the hypothesis tests often find no significant evidence to reject either null hypothesis. As expected, the percentage of times no significant evidence was found to reject either null hypothesis increases as $\alpha \rightarrow 1$ and $\beta \rightarrow 0$. These patterns are seen in both the stochastic polynomial and logarithmic maps and the stochastic polynomial-logarithmic map.

Due to the higher penalty term, the use of BIC chose a single parameter map more often than the use of AIC. When the true generating map was either the stochastic polynomial map or the stochastic logarithmic map, this led to BIC outperforming AIC. However, when the true generating map was a stochastic polynomial-logarithmic map the opposite is true. Both criteria selected the same map for the majority of the orbits.

As with the use of the hypothesis tests, the correct model was less likely to be identified by AIC and BIC as $\alpha \rightarrow 1$ and/or $\beta \rightarrow 0$. The results for the stochastic polynomial-logarithmic map, in particular, show a smaller value of α increases the percentage of times a stochastic polynomial map is fitted, whilst a larger value of β increases the percentage of times a stochastic logarithmic map is fitted.

Table 6.5 shows the percentages the AIC, BIC and hypothesis testing methods chose each map for each group of orbits with $n = 1000$ and $\theta = 1.5$. The percentage of times in which the correct model is identified is much higher for all three methods of model selection than the results for $n = 100$. For each case other than the stochastic polynomial map with $\alpha = 0.9$ the correct map is identified more than 59% of the time, and usually more than 80%. The patterns picked up from the results for $n = 100$ still appear present, and the correct model appears less likely to be identified as $\alpha \rightarrow 1$ and/or $\beta \rightarrow 0$.

For the stochastic polynomial and logarithmic maps the use of BIC outperforms the use of AIC or the hypothesis tests in identifying the correct model. However, when the true generating map is the stochastic polynomial-logarithmic map, the use of BIC identifies the correct map the least out of the three selection methods.

The only situation in which the correct map is not identified the majority of the time is for the stochastic polynomial map with $\alpha = 0.9$ using the hypothesis tests, for which the non-stationary model was not rejected for 80% of the orbits. This is because the theoretical standard deviation of the estimate of α from equation 6.16 is approximately 0.08, hence at the 5% significance level the null hypothesis of $\alpha = 1$ will only be rejected for estimates of $\alpha < 0.84$, which should theoretically occur only 23% of the time. This sample frequency of 20% is thus in near agreement with the theoretical results.

$n = 1000, \theta = 1.5$										
Polynomial Map	AIC			BIC			Hypothesis test			
	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.3$	96	0	4	100	0	0	98	0	2	0
$\alpha = 0.45$	91	0	9	99	1	0	94	0	6	0
$\alpha = 0.5$	85	0	15	98	0	2	90	0	10	0
$\alpha = 0.65$	93	1	6	98	2	0	96	1	1	2
$\alpha = 0.8$	83	15	2	85	15	0	67	3	2	28
$\alpha = 0.9$	60	33	7	67	33	0	18	0	2	80
Logarithmic Map	α	β	Both	α	β	Both	α	β	Both	None
$\beta = 0.05$	0	92	8	1	99	0	0	96	4	0
$\beta = 0.15$	0	97	3	0	99	1	0	97	3	0
$\beta = 0.25$	0	94	6	0	100	0	0	96	4	0
$\beta = 0.3$	0	90	10	0	100	0	0	95	5	0
Poly-Log Map	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.45, \beta = 0.05$	0	0	100	0	0	100	0	0	100	0
$\alpha = 0.45, \beta = 0.3$	0	0	100	0	0	100	0	0	100	0
$\alpha = 0.8, \beta = 0.05$	0	13	87	0	41	59	0	15	85	0
$\alpha = 0.8, \beta = 0.3$	1	14	85	1	38	61	1	15	84	0

Tab. 6.5: the percentages the AIC, BIC and hypothesis testing methods chose each map for each group of orbits with $n = 1000$ and $\theta = 1.5$. The headings α , β , 'Both' and 'None' refer to the method choosing the stochastic polynomial map, the stochastic logarithmic map, the stochastic polynomial-logarithmic map and the non-stationary stochastic polynomial-logarithmic map respectively.

$n = 4000, \theta = 1.5$										
	AIC			BIC			Hypothesis test			
Polynomial Map	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.3$	90	0	10	100	0	0	93	0	7	0
$\alpha = 0.45$	91	0	9	100	0	0	96	0	4	0
$\alpha = 0.5$	86	0	14	100	0	0	94	0	6	0
$\alpha = 0.65$	93	0	7	100	0	0	97	0	3	0
$\alpha = 0.8$	92	0	8	100	0	0	97	0	3	0
$\alpha = 0.9$	79	14	7	86	14	0	75	3	3	19
Logarithmic Map	α	β	Both	α	β	Both	α	β	Both	None
$\beta = 0.05$	0	90	10	0	100	0	0	94	6	0
$\beta = 0.15$	0	94	6	0	100	0	0	98	2	0
$\beta = 0.25$	0	90	10	0	100	0	0	91	9	0
$\beta = 0.3$	0	93	7	0	100	0	0	97	3	0
Poly-Log Map	α	β	Both	α	β	Both	α	β	Both	None
$\alpha = 0.45, \beta = 0.05$	0	0	100	0	0	100	0	0	100	0
$\alpha = 0.45, \beta = 0.3$	0	0	100	0	0	100	0	0	100	0
$\alpha = 0.8, \beta = 0.05$	0	0	100	0	0	100	0	0	100	0
$\alpha = 0.8, \beta = 0.3$	0	0	100	0	0	100	0	0	100	0

Tab. 6.6: the percentages the AIC, BIC and hypothesis testing methods chose each map for each group of orbits with $n = 4000$ and $\theta = 1.5$. The headings α , β , 'Both' and 'None' refer to the method choosing the stochastic polynomial map, the stochastic logarithmic map, the stochastic polynomial-logarithmic map and the non-stationary stochastic polynomial-logarithmic map respectively.

The results for the stochastic polynomial map with $\alpha = 0.9$ also select the correct map a much lower percentage of the time than the other values of α using AIC and BIC. This is expected, since the stochastic logarithmic map has a value of $\alpha = 1$, and thus discrimination between the maps when $\alpha \rightarrow 1$ becomes more difficult.

Table 6.6 shows the corresponding percentages for $n = 4000$. For this length of observations, the correct map is identified at least 75% of the time in all cases and generally more than 90% of the time. The pattern of the frequency of correctly identified map decreasing as $\alpha \rightarrow 1$ and/or $\beta \rightarrow 0$ is less clear than with the smaller series lengths, although the results for the stochastic polynomial map with $\alpha = 0.9$ are still significantly worse than the results for the other values of α for all the methods. As previously mentioned, this is due to the stochastic logarithmic map having a value of $\alpha = 1$, and thus there being less difference between the maps when $\alpha = 0.9$.

The use of BIC does particularly well, with the correct model identified 100% of the time in each case other than the stochastic polynomial map with $\alpha = 0.9$ for which 14% of the orbits are mistakenly identified as generated from a stochastic logarithmic map.

Table 6.7 shows the effect of θ on the percentage of times the correct model was identified using each selection method for $n = 100$. It can be seen that as θ decreases the percentage of correctly identified models increases. For $\theta = 0.5$ the correct map was identified in the majority of cases even for $n = 100$. The previously noted patterns that were present for $\theta = 1.5$ still appear for $\theta \in (0.5, 1)$. The use of AIC and BIC still seem to outperform the use of hypothesis tests for the stochastic polynomial and logarithmic maps, whilst BIC gives the worst results for the stochastic polynomial-logarithmic map. The correct model appears less likely to be identified as $\alpha \rightarrow 1$ and/or $\beta \rightarrow 0$ for each value of θ .

The effect of θ for $n \in (1000, 4000)$ is similar to that for $n = 100$ in that as θ decreases the percentage of correctly identified models generally increases. However, even for $\theta = 1.5$ all three selection methods tend to choose the correct model, as seen in tables 6.5 and 6.6, thus the effect of θ is generally far less noticeable and the tables of results are thus omitted. The most significant change caused by θ is for the stochastic polynomial map with $\alpha = 0.9$, with the percentage of times the correct map is identified by the hypothesis testing increasing from 18% to 85% for $n = 1000$ and $\theta = 0.5$ and increasing from 75% to 95% for $n = 4000$ and $\theta = 0.5$.

The results of this simulation study show that due to the large variance of the estimates of α and β when estimated simultaneously, the use of a criterion such as AIC or BIC is often preferred to hypothesis testing to correctly identify the generating map when it is assumed unknown. As the orbit length increases, the probability of correctly identifying the generating map also increases and for $n \geq 1000$ the fitted map is generally reliable for $\alpha \leq 0.8$ and/or $\beta \geq 0.05$. Having found the correct map, the results from section 6.4.2 suggest consistent estimates of α and β can be found and the results of section 6.4.1 suggest these are equivalent to

estimating the asymptotic values of the long memory parameters d and c .

$n = 100$									
θ	AIC			BIC			Hypothesis test		
	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Polynomial Map									
$\alpha = 0.3$	94	89	81	99	92	85	94	71	55
$\alpha = 0.45$	83	89	74	92	91	76	83	66	27
$\alpha = 0.5$	87	82	73	91	86	77	82	52	32
$\alpha = 0.65$	83	64	57	89	66	61	64	19	9
$\alpha = 0.8$	62	57	54	67	61	56	28	7	3
$\alpha = 0.9$	67	45	50	71	46	54	8	5	1
Logarithmic Map									
$\beta = 0.05$	91	70	67	98	76	67	89	39	15
$\beta = 0.15$	89	83	70	97	85	72	90	47	31
$\beta = 0.25$	94	80	69	99	82	73	94	55	32
$\beta = 0.3$	92	84	73	98	87	75	96	64	35
Poly-Log Map									
$\alpha = 0.45, \beta = 0.05$	97	50	14	86	17	4	93	34	13
$\alpha = 0.45, \beta = 0.3$	97	61	21	94	30	9	97	45	13
$\alpha = 0.8, \beta = 0.05$	49	3	2	22	0	0	37	0	1
$\alpha = 0.8, \beta = 0.3$	50	8	5	31	2	0	42	3	1

Tab. 6.7: The percentages the AIC, BIC and hypothesis testing selection methods chose the correct map for each group of orbits with $n = 100$ and $\theta \in (0.5, 1, 1.5)$.

Part II

7. DEODORANT STICK DATA ANALYSIS

The second part of this thesis focuses on the analysis of consumer data provided by Unilever. The data is concerned with the movement of individuals over time and space whilst applying a deodorant stick to the area under their left arm, the raw data being the (x, y, z) co-ordinates with time stamp of seven sensors attached to the individual.

The deodorant stick data being discussed in this paper has been collected and supplied by Unilever and is part of a larger series of experiments being carried out using the company's recently acquired motion sensor technology. Motion sensor technology has been used previously in several areas, such as virtual simulations of factories and equipment to test and improve designs, see for example Faraway and Reed [2007]. There is also increasing demand for such technology in entertainment, such as movies and video games, as well as fields such as sport and medicine, see Menache [2000]. In addition to the deodorant stick data studied here, Unilever is also starting to apply motion sensors to other product areas such as brushing teeth and combing hair.

Unilever has several objectives in carrying out this study. First, it would be of interest to them to discover if such an experiment can pick up differences between individuals and products. Assuming such differences could be found, it would then be useful to group individuals and products into different application techniques and categories. Also, the cost of various experiments at Unilever could be greatly reduced if suitable methods of modelling the data and simulating new data could be found. Unilever has recently obtained robotic arms which could be used to replace many experiments on individuals at a reduced cost provided that the simulated data is realistic. More generally, the use of such sensors is still in its early stages for experiments at Unilever and they are being used in an increasingly large range of applications. Therefore, it would be of great use to Unilever in terms of designing future experiments to see examples of what methods of analysis can be applied.

A description of the deodorant stick data, including how it was collected and recorded is given in section 7.1. Section 7.2 looks at a univariate series from the data which Unilever team believed may give useful information about the differences between individuals and products during application. To this aim, ANOVA techniques are introduced and applied in sections 7.2.2 - 7.2.3.

Section 7.3 attempts to better interpret the data. After principal component analysis is carried out on the data in section 7.3.2, a new form of the data is derived in section 7.3.3. This new transformed form of the data provides a reduction in

dimensionality without loss of information whilst simultaneously providing easier to understand definitions to the series.

Finally, section 7.4 looks at methods of modelling the data in this new form and simulating new data with the same properties as the original. Section 7.4.2 introduces the use of vector autoregressive models, whilst section 7.4.3 looks instead at modelling the data by fitting functions such as B-Splines and Bezier curves.

7.1 Data Collection

The study of the deodorant stick data involves ten different individuals and four different deodorant stick products. Details of these individuals and products have not been supplied and thus they will be referred to only as Person i , $i \in (1, \dots, 10)$ and Product j , $j \in (1, \dots, 4)$. The order of these labels was chosen arbitrarily. The majority of (i, j) combinations were repeated 6 times, although four were repeated 7 times and Person 8 used each product only once, see table 7.1. Each set of data is thus labelled as X_{ijk} , with $k \in (1, \dots, 7)$ representing the repetition number and i and j representing the Person and Product numbers respectively. There are a total of 224 sets of data.

Product (j)	Person (i)									
	1	2	3	4	5	6	7	8	9	10
1	7	6	6	6	6	6	6	1	6	6
2	6	6	6	6	6	6	6	1	6	6
3	6	6	6	6	6	6	6	1	6	7
4	7	6	6	6	6	6	6	1	6	7

Tab. 7.1: Number of repetitions, K , for each (i, j) combination.

For each (i, j, k) th trial, Person i applied Product j to their left armpit region, whilst the (x, y, z) co-ordinates of seven sensors were recorded at equally spaced time intervals, with 100 observations per second. Each individual was not told when to start and stop and thus the application time varies from 1.05 seconds to 15.59 seconds, with a mean of 5.23 seconds and a standard deviation of 3.50 seconds. The majority of data sets also have a few seconds of recorded observations before and after application, the lengths of which also vary greatly.

The seven sensors were attached to the following seven locations for each (i, j, k) th trial:

1. The head of the deodorant stick of Product j .
2. The bottom of the deodorant stick of Product j .
3. The left wrist of Person i .
4. The left elbow of Person i .

5. A point on the left side of Person i , approximately 30cm directly below the left shoulder.

6. A point on the left side of Person i , approximately 40cm directly below the left shoulder.

7. The left shoulder of Person i .

For the remainder of this paper the sensors will be numbered according to this list, the order of which was decided originally by Unilever. For each (i, j, k) th trial, there are 21 time series, the (x, y, z) co-ordinates producing three for each sensor. Thus each $X_{ijk}(t)$, the observations from X_{ijk} at a time t , is a vector of length 21.

Sensors 1 and 2 were fitted to describe the movement of the deodorant stick during application. Sensors 3, 4 and 7 are assumed to give information on the location and movement of the left arm which is lifted during the application process. Sensors 5 and 6 were fitted in an attempt to record the effect of the application process to the underarm of the individual. In particular, it was assumed that products with greater friction would cause the underarm skin to stretch and contract more during application and that differences in products may therefore be detected by the distances between these two sensors. This claim is studied further in section 7.2

The results were intended to be as natural as possible and to this aim the individuals were also not told an exact place to stand nor an exact direction to face. The result of this is that the (x, y, z) co-ordinates are not all recorded from the same x, y and z axis relative to the individual. Indeed, each individual was free to move around even during application if so desired and thus the axes may not be fixed relative to the individual for even the duration of one trial. In order to avoid these difficulties, the use of a new set of data, Y_{ijk} , was introduced, which records the 21 Euclidean distances between each pair of sensors, using the standard formula for finding the distance at each point in time, d , between two points (x_a, y_a, z_a) and (x_b, y_b, z_b) as

$$d = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}.$$

For ease, the notation $\{a, b\}_{ijk}$ will be used to represent the time series generated by the distances between sensors a and b for the ijk th trial, where the sensors are numbered by the list given above. Also, the subtext ijk may be dropped when not necessary and it will be made clear from the context if referring to a specific trial.

The study of this multivariate data Y_{ijk} begins in section 7.3. In section 7.2 an initial study of the univariate data obtained from the distance between sensors 5 and 6, series $\{5,6\}$, is carried out.

It should also be noted that the 224 trials were taken on several different days over several months. However, early studies by Unilever have suggested no significant reason to suggest the day of the trial will have an effect on the data, and this variate has thus been ignored for simplicity.

7.2 Under Arm Data

The original idea for placing two sensors on the side of the individual under the left arm (sensors 5 and 6 from the list in section 7.1) was the belief by the Unilever team that the stretching and contracting of the skin between these two sensors would give useful information about the differences between individuals and products during application. The time series defined by the distance between these two sensors over time is therefore studied first as a univariate case to see what information could be obtained.

Since the start and end times of the deodorant stick application are unknown they must first be estimated from the data. This is discussed in section 7.2.1, which introduces change-point analysis to determine these points. Analysis of variance techniques are then applied to various statistics from the series in section 7.2.2 to investigate if product and individual effects are present. Section 7.2.3 continues this ANOVA in the frequency domain.

7.2.1 Change Point Analysis

For each of the 224 trials, the underarm series, {5,6}, follows a general pattern. It begins fairly constant whilst the individual is at rest, before application. There is then a sudden sharp increase as the individual raises their left arm and the skin between sensors 5 and 6 stretches. The distance between the sensors remains at this higher level during application of the deodorant stick, stretching and contracting a relatively small amount due to the movement of the deodorant stick. When the application is finished, there is a sudden sharp decrease as the individual lowers their left arm, followed by a fairly constant level as the individual returns to rest. An example can be seen in figure 7.1.

Since the deodorant stick application process is of interest in the study, it would be useful to be able to take out the 'middle section' of each time series, namely the section after the arm is raised until the arm is lowered, during which time the application occurs. This can be done quite easily by inspection after plotting the data. However, inspection of 224 time series can be a lengthy task and subject to opinion. It was therefore useful to develop an algorithmic method of deciding on the start and end times of this middle section.

The problem of finding unknown points such as the start and end points of the application can be approached by change-point analysis. Early work on this subject can be found in Page [1955], [1957] and various others such as Pettitt [1980], Hinkley [1971] and Hinkley and Schechtman [1987] etc. Various tools are suggested in these papers to help find unknown change-points and indeed to discover if there is any significant evidence to suggest a change point is present at all. Since the existence of the change points are known, proving the points to be significant is unnecessary and instead focus is given only on finding them. Two examples of simple approaches of locating a change point are the CUSUM plot and a mean squared error (MSE)

approach.

Let x_t , $t \in (1, \dots, T)$, be an observed time series, with estimated mean \bar{x} . Then the CUSUM plot is a plot of S_i against i , $i \in (1, \dots, T)$, where

$$S_i = \sum_{t=1}^i x_t - \bar{x}. \quad (7.1)$$

Now, it is assumed that x_t is of the form

$$x_t = \mu_1 + \varepsilon_t, \quad t < m, \quad x_t = \mu_2 + \varepsilon_t, \quad t > m \quad (7.2)$$

and thus a change has occurred at the unknown point $t = m$. From equation 7.1 it can be seen that a positive gradient in the CUSUM plot is produced by values above the mean and a negative gradient is produced by those below. Long sections of positive or negative gradient would thus be produced by a large series of consecutive observations above or below the mean. This suggests a change-point in mean most likely to have occurred at the max or min of the CUSUM plot, giving the estimated change point, \hat{m} , as

$$\hat{m} = \arg \max |S_m|$$

The MSE approach also assumes x_t to be of the form given in equation 7.2 and thus gives the estimate of m as

$$\hat{m} = \arg \min \sum_1^m (x_t - \bar{x}_{m1})^2 + \sum_{m+1}^T (x_t - \bar{x}_{m2})^2$$

where

$$\bar{x}_{m1} = \frac{\sum_1^m x_t}{m}, \quad \bar{x}_{m2} = \frac{\sum_{m+1}^T x_t}{T - m}.$$

For each method, once one change point is found, the observed time series can be split and the further change points can be found by repeating the method on the sample up to or from the first change point.

However, for the deodorant stick data these basic methods are not suitable in their present form. They tend to split the data such that the estimated starting point is too early and the estimated end point is too late. The reason for this is that the time series do not have an instantaneous change in mean at a single point but a sudden steep trend, changing the mean gradually over several points. The assumption of x_t following equation 7.2 therefore does not hold. In fact, there are four change points to be found, separating the sections into a relatively zero trend, then a strong positive trend, then a relatively zero trend, a strong negative trend

and a final flat trend. A suitable extension of the MSE approach for the deodorant stick data can be given under the new assumption that x_t is of the form

$$x_t = \begin{cases} \mu_1 + \varepsilon_t & 1 \leq t < m_1 \\ \mu_2 + b_1 t + \varepsilon_t & m_1 \leq t < m_2 \\ \mu_3 + \varepsilon_t & m_2 \leq t < m_3 \\ \mu_4 - b_2 t + \varepsilon_t & m_3 \leq t < m_4 \\ \mu_5 + \varepsilon_t & m_4 \leq t < T \end{cases}, \quad (7.3)$$

where $1 < m_i < m_{i+1} < T$, the μ_i for $i \in (1, \dots, 5)$ are constants and $b_1, b_2 > 0$ represent the trends created whilst raising and lowering the arm and ε_t is the error term. The application process thus takes place when $t \in (m_2, \dots, m_3)$. For a given set of change points, (m_1, m_2, m_3, m_4) , the model given in equation 7.3 can be fitted in a piecewise manner using least squares estimates of μ_i for $i \in (1, \dots, 5)$ and b_1, b_2 . The residuals from this fitted model, $\hat{\varepsilon}_t$ $t \in (1, \dots, T)$, are thus dependent on the change points (m_1, m_2, m_3, m_4) and the extended MSE approach of finding these change points is to use the estimates which minimise the mean of squares of these residuals,

$$(\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{m}_4) = \arg \min \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t(m_1, m_2, m_3, m_4)^2. \quad (7.4)$$

The computational time of finding the MSE for all possible combinations of (m_1, m_2, m_3, m_4) can be considerable. This time can be greatly reduced by restricting the search area for each change point. Using the CUSUM plot, the local min of S_i will be a point between m_1 and m_2 and the local max will be a point between m_2 and m_3 , since the shape of x_t is such that the overall mean, μ , will be $\mu_1 < \mu < \mu_3$ and $\mu_5 < \mu < \mu_3$. Let the local min and max of the CUSUM be c_1 and c_2 respectively. Now, it is known from studying the deodorant stick data that the raising and lowering of the arm generally takes around 1 second, that is $m_2 - m_1 \approx 100$ and $m_4 - m_3 \approx 100$, hence it is assumed that $m_1 \in (c_1 - 100, c_1)$, $m_2 \in (c_1, c_1 + 100)$, $m_3 \in (c_2 - 100, c_2)$ and $m_4 \in (c_2, c_2 + 100)$. This has reduced the number of possible combinations of (m_1, m_2, m_3, m_4) from $O(T^4)$ to $100^4 = 10^8$.

The computational time can be reduced further by working in steps, minimising equation 7.4 first for every 20th value of each of (m_1, m_2, m_3, m_4) , to find initial estimates $(\hat{m}_{11}, \hat{m}_{21}, \hat{m}_{31}, \hat{m}_{41})$. The search areas are then reduced to assuming $m_1 \in (\hat{m}_{11} - 10, \hat{m}_{11} + 10)$, $m_2 \in (\hat{m}_{21} - 10, \hat{m}_{21} + 10)$, $m_3 \in (\hat{m}_{31} - 10, \hat{m}_{31} + 10)$ and $m_4 \in (\hat{m}_{41} - 10, \hat{m}_{41} + 10)$. These areas are reduced again by repeating this process looking at every 5th value and the finally every value is considered. This reduces the number of (m_1, m_2, m_3, m_4) combinations considered from 10^8 to $3 \times 5^4 = 1875$.

A brief simulation study was carried out to test this method of finding the change

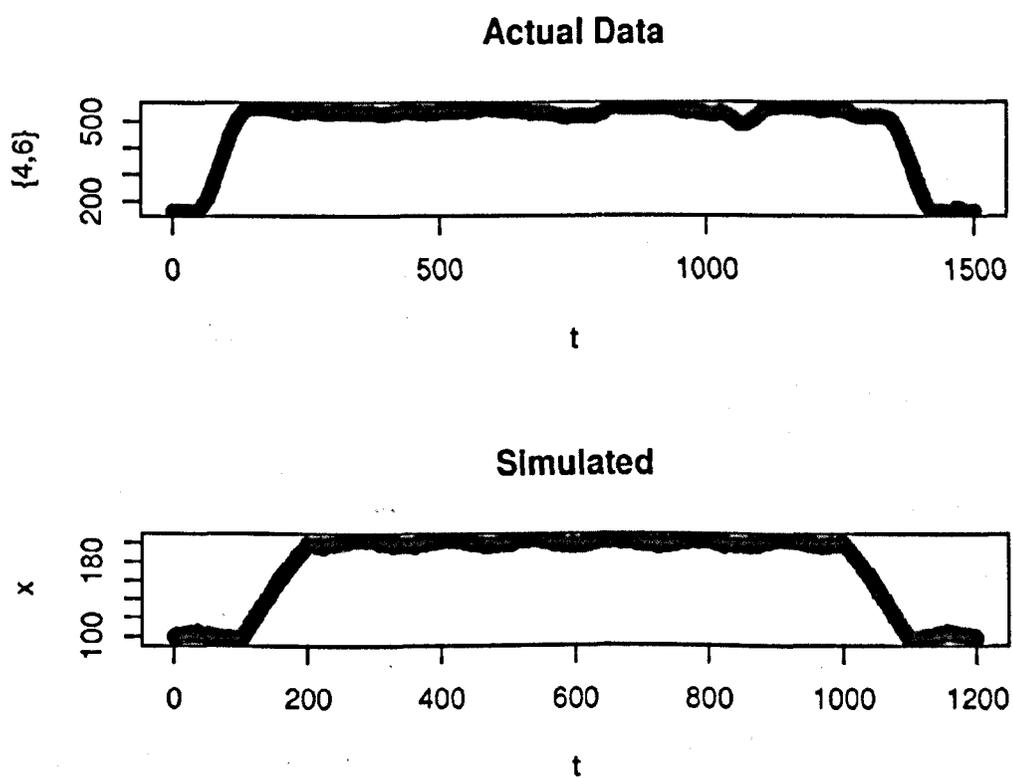


Fig. 7.1: Examples of the shape of actual and simulated data from which change points were estimated.

points (m_1, m_2, m_3, m_4) . The series x_t was simulated with $m_1 = 100$, $m_2 = 200$, $m_3 = 1000$, $m_4 = 1100$, $T = 1200$, $\mu_1, \mu_5 = 100$, $\mu_3 = 200$, $\mu_2 = 0$, $\mu_4 = -800$, $b_1, b_2 = 1$ and

$$\varepsilon_t = 3 \sin\left(\frac{t}{20}\right) + \eta_t,$$

where $\eta_t \sim N(0, 1)$ are IID. The choice of these values are somewhat arbitrary, chosen simply to ensure the overall general appearance of the simulated data is similar to that of the actual data. Figure 7.1 shows an example of an actual time series and a simulated series for comparison.

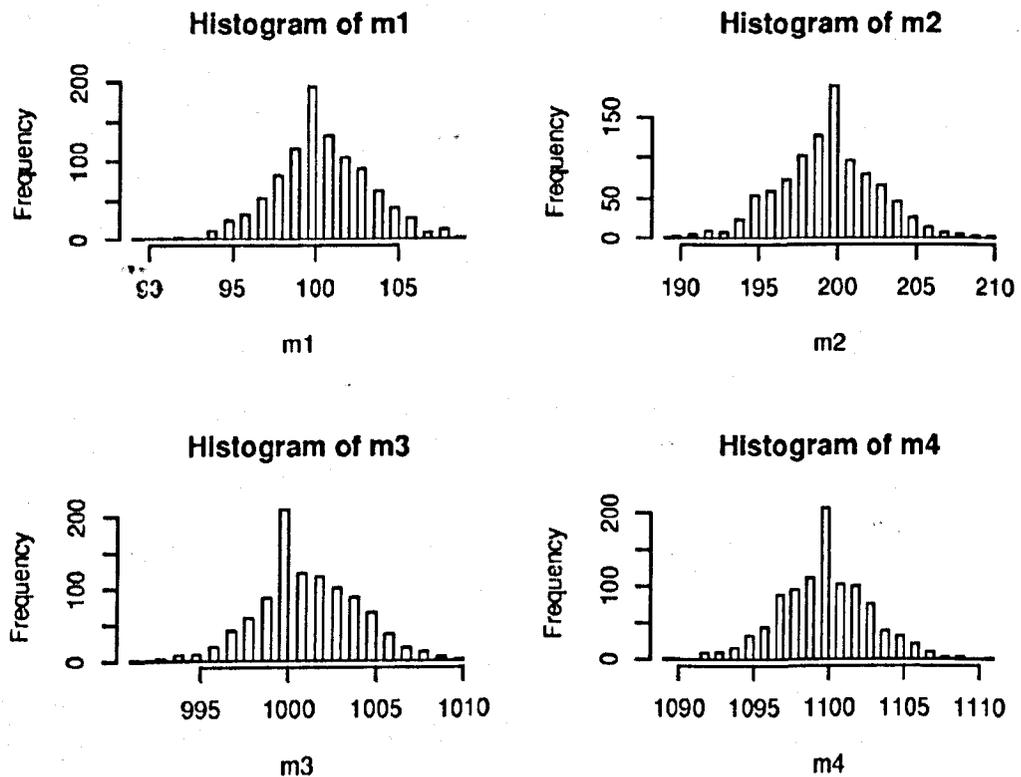


Fig. 7.2: Histograms of the estimated values of (m_1, m_2, m_3, m_4) from 1000 simulations of x_t .

Estimates of the change points (m_1, m_2, m_3, m_4) were found and recorded for x_t using the previously suggested method. The simulation was repeated 1000 times, with new η_t being generated each time and the other parameters remaining constant and the values of $(\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{m}_4)$ recorded for each. The histograms of these estimates are given in figure 7.2.

It can be seen that the true values of (m_1, m_2, m_3, m_4) are found approximately 20% of the time, whilst the vast majority of the estimates are within ± 10 of the true values. Assuming this to be an accurate reflect of the deodorant stick data suggests the application start and end times can be estimated to within 0.1 sec.

When estimating the change points from the actual data, it was decided to use the series $\{4, 6\}$ rather than $\{5, 6\}$, that is, the distance between the elbow sensor and the lower underarm sensor. This clearly must follow the same general pattern as $\{5, 6\}$, increasing as the arm is raised and decreasing as the arm is lowered, hence the points (m_1, m_2, m_3, m_4) are the same for both, but the series $\{4, 6\}$ was chosen as it was generally the series which most closely matched the assumed form of x_t , being less effected by the movement of the deodorant stick.

7.2.2 ANOVA Application

An application of ANOVA techniques to the underarm data, series $\{5, 6\}$, is carried out here by summarising each of the 224 time series into several easy to interpret statistics and carrying out an ANOVA on each. The change point analysis is first carried out on the series $\{4, 6\}$ to find the change points (m_1, m_2, m_3, m_4) for each series. These points are then used on the series $\{5, 6\}$. The middle section between points m_2 and m_3 is considered to represent the application time. The time from $t = 1$ to m_1 is referred to as the 'before' section, whilst the time from m_4 to T is referred to as the 'after' section of the series.

1. Middle Section Mean and variance:- It was believed that the distance between the two sensors of the underarm during application could give an idea of how much the deodorant stick pulls and stretches the skin during application. It may be logical, therefore, to assume greater friction may cause more stretching and contracting, resulting in a possible difference in mean and/or variance. If a significant difference is found, it may be possible to determine which product creates greatest friction with the individual.

2. Middle Section Starting Point, Ending Point and Length:- While applying the deodorant, each individual was not told exactly when to start or finish the application. The individual was only told when the recording of the data had begun and were then allowed to lift their arm, apply the deodorant and lower their arm in their own time. This puts forward the possibility that the length of application may subconsciously indicate a preference to a certain product. i.e. The more a person likes the feel of a certain product, the longer they may wish to apply it and vice versa. The starting and end points refer to the points m_2 and m_3 respectively. Assuming the person has no prior knowledge of the products, the starting points would be expected to show no dependence on the product and may show nothing more than the reaction time of the individual. The products were all supplied in identical cases, so no opinion could be made of the products, subconscious or otherwise, via appearance. If the starting points did show signs of significant difference between products, it could possibly be accounted for by smell. The end points, assuming fairly

constant starting points would give results similar to the analysis of the lengths of the middle sections.

3. Before and After sections mean and variance:- As with the motivation behind the study of the lengths of the middle sections, it was suggested that a study of the 'end sections' may also show subconscious behaviour, 'end sections' here referring to the section of the series with $t \in (1, m_1)$ and $t \in (m_4, T)$. Both sections may be considered to show the person at rest. However, if a product feels 'sticky', the individual may not lower their arm as much, which should result in an increased mean for the end section. They may also move their arm slightly more due to its less comfortable position resulting in a higher variance. Thus the difference between the mean and variance of the first and end section, i.e. the increase of the mean and variance due to the 'sticky' effect was also recorded.

4. Overall mean and variance:- The overall mean and variance, i.e. the mean and variance of all available observations for each set of underarm data, would likely be most strongly influenced by the raising of the arm. It is therefore conceivable that a difference in feel of the deodorant stick products may cause the individual to raise their arm more or less which could provide information about the products.

For each of these statistics, h_{ijk} say, one-way and two way ANOVA techniques were carried out. For the two way ANOVA results, one set of data was removed from each (i, j) combination with 7 repetitions to give proportional sampling and thus an orthogonal design matrix. The tests were carried out for each statistic based on the three models given below in equations 7.5, 7.6 and 7.7.

$$h_{ijk} = \mu + A_i + \varepsilon_{ijk}, \quad (7.5)$$

$$h_{ijk} = \mu + B_j + \varepsilon_{ijk} \quad (7.6)$$

and

$$h_{ijk} = \mu + A_i + B_j + C_{ij} + \varepsilon_{ijk}, \quad (7.7)$$

where A_i is the effect of the i th Person on the (i, j, k) th statistic, B_j is the effect of the j th Product on the (i, j, k) th statistic, C_{ij} is the interaction effect of the i th Person and j th Product on the (i, j, k) th statistic and ε_{ijk} are assumed to be independent and normally distributed. The results are given in table 7.2.

It must be noted that the Gaussian assumption on the errors may not hold and thus the reliability of these results is questionable. The Normal QQ plots from the residuals of the middle section means are shown in figure 7.3. Figure 7.3a) gives the QQ plot of the residuals under the assumption of a constant mean regardless of individual or product, and figures 7.3b)-d) give the QQ plot of the residuals of equations 7.5-7.7 respectively. The QQ plots for the other statistics are similar in appearance and thus omitted.

It can be seen that there is some evidence at the 5% level to reject the null hypothesis that the residuals are normally distributed near the upper tails. However,

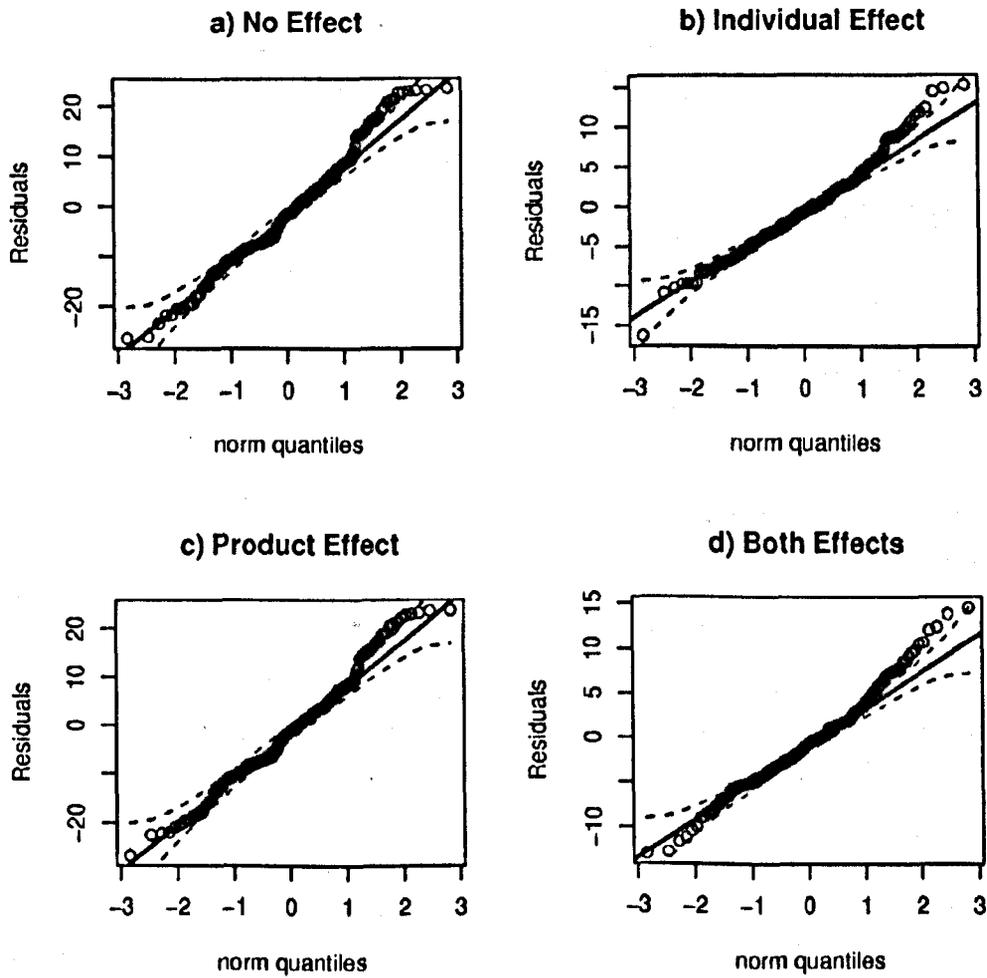


Fig. 7.3: Normal QQ plots of the residuals of the models fitted by equations 7.5-7.7 from the middle section means. The dotted lines represent the 95% confidence intervals that the residuals are normally distributed.

Statistic	Model 1	Model 2	Model 3		
	Person	Product	Person	Product	Interaction
Middle Section Mean	***		***		
Middle Section Variance					
Middle Section Length	***		***		
Starting Point	***		***		
End Point	***		***		
Before Section Mean	***		***		
Before Section Variance					
After Section Mean	***		***		
After Section Variance					
Difference in before/after Means	**				
Difference in before/after Variances					
Overall Mean	***		***		
Overall Variance	***		***		

Tab. 7.2: ANOVA results for the statistics of the Under Arm Data. * - significant at 5% level, ** - significant at 1% level, *** - significant at 0.1% level. A blank box shows no significance at the 5% level.

this deviation from the normal distribution seems slight, whilst the significant results are generally significant even at the 0.1% level. There appears to be strong significant evidence to suggest that different individuals may have an effect on the majority of these statistics for the data. It suggests that different individuals possess different application techniques for the deodorant sticks.

In comparison, no significant evidence is seen to suggest any difference between products. The current explanation suggested by Unilever is that the individuals may subconsciously alter the force with which they apply the deodorant sticks such that they always achieve the amount of friction they are used to. If this is the case it may be very difficult to find any way to distinguish between products from this data. Future experiments of Unilever are intended to record also the amount of force applied during application, which would be useful in further investigating this claim.

A study of human motion during an experiment done by Faraway et al [1999], involving reaching, showed motion can be dependent on many factors concerning the individual, such as age, gender, joint mobility and muscle strength. Such information is not available for the data studied here, though future studies may wish to explore further which factors are significant for differences in deodorant stick application.

7.2.3 Frequency Domain ANOVA Analysis

Having carried out an ANOVA for statistics in the time domain, attention now turns to the frequency domain. With the focus still to discover any significant common components between different people or products, the frequency analysis of time series collected in an experimental design suggested by Brillinger [1972] is carried

out. Since a common method of deodorant stick application may consist of a repeated up and down motion, a periodic or near periodic element may be present, making frequency analysis particularly appealing.

For simplicity, it is first assumed that only a common component in products may be present, that is

$$x_{ijk}(t) = \mu_{ijk} + a(t) + b_i(t) + \varepsilon_{ijk}(t) \quad (7.8)$$

with $i \in (1, \dots, I)$ representing the person number, $j \in (1, \dots, J)$ representing the product number and $k \in (1, \dots, K)$ representing the repetition number and where μ_{ijk} are constants, $a(t)$ is a stationary time series common to all x_{ijk} , the $b_i(t)$ are stationary time series common to each person and the $\varepsilon_{ijk}(t)$ are stationary time series unique to each x_{ijk} . It is assumed that the cumulants of $a(t), b_i(t)$ and $\varepsilon_{ijk}(t)$ are all absolutely sumable and the means of each series is zero. In this first model it is assumed that no effect of different products is present and thus these time series can be taken as further replications. A model including possible effects from both individuals and products is discussed later.

Looking at the model given in equation 7.8, it follows that

$$E(x_{ik}) = \mu_{ik}$$

$$f_{x_{ik}x_{ik}}(\lambda) = f_{aa}(\lambda) + f_{bb}(\lambda) + f_{\varepsilon\varepsilon}(\lambda)$$

$$f_{x_{ik_1}x_{ik_2}}(\lambda) = f_{aa}(\lambda) + f_{bb}(\lambda), \quad k_1 \neq k_2 \quad (7.9)$$

$$f_{x_{i_1k_1}x_{i_2k_2}}(\lambda) = f_{aa}(\lambda), \quad i_1 \neq i_2 \quad (7.10)$$

where, since no product effect is assumed, the subscript j has been removed and now $k \in (1, \dots, JK)$ is the repetition number. Here f_{yz} represents the power spectrum given by

$$f_{yz}(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \text{cov}(y_t, z_{t-u}) e^{-i\lambda u}.$$

From equation 7.9 and equation 7.10 it can be seen that if $f_{bb}(\lambda) = 0$, time series from the same individual will have no greater linear dependency than time series from different individuals and if $f_{aa}(\lambda) = 0$ time series from two different individuals will not be linearly dependent. Brillinger [1972] points out that given x_{ik} is of the form in equation 7.8, it can be written that

$$X_{ikl} = A_l + B_{il} + E_{ikl} \quad (7.11)$$

$l \in (1, \dots, L)$, where X_{ikl} , A_l , B_{il} and E_{ikl} are the discrete Fourier transforms of $x_{ik}(t)$, $a(t)$, $b_i(t)$ and $\varepsilon_{ik}(t)$ at λ_l , respectively. The advantage of this notation is that X_{ikl} is now asymptotically complex normally distributed with mean 0 and covariance

matrix $f_{xx}(\lambda)$ and A_l, B_{il} and E_{ikl} are asymptotically complex normally distributed with mean 0 and covariance matrixes $f_{aa}(\lambda), f_{bb}(\lambda)$ and $f_{ee}(\lambda)$ respectively. Applying standard ANOVA techniques to equation 7.11 then gives two F statistics for each value of λ to test the significance of $f_{aa}(\lambda)$ and $f_{bb}(\lambda)$. The first, under the null hypothesis that $f_{aa}(\lambda) = 0$, is given by

$$\frac{IK \hat{f}_{\bar{x}.., \bar{x}..}(\lambda)}{\sum_i \sum_k \hat{f}_{x_{ik} - \bar{x}_i, x_{ik} - \bar{x}_i}(\lambda) / (I(K-1))} \sim F_{2L, 2I(K-1)L}$$

The second, under the null hypothesis that $f_{bb}(\lambda) = 0$, is given by

$$\frac{K \sum_i \hat{f}_{\bar{x}_i - \bar{x}., \bar{x}_i - \bar{x}..}(\lambda) / (I-1)}{\sum_i \sum_k \hat{f}_{x_{ik} - \bar{x}_i, x_{ik} - \bar{x}_i}(\lambda) / (I(K-1))} \sim F_{2(I-1)L, 2I(K-1)L}$$

where $\hat{f}_{\bar{x}.., \bar{x}..}(\lambda), \hat{f}_{x_{ik} - \bar{x}_i, x_{ik} - \bar{x}_i}(\lambda)$ and $\hat{f}_{\bar{x}_i - \bar{x}., \bar{x}_i - \bar{x}..}(\lambda)$ are the sample estimates of $f_{\bar{x}.., \bar{x}..}(\lambda), f_{x_{ik} - \bar{x}_i, x_{ik} - \bar{x}_i}(\lambda)$ and $f_{\bar{x}_i - \bar{x}., \bar{x}_i - \bar{x}..}(\lambda)$ given by

$$\hat{f}_{\bar{x}.., \bar{x}..}(\lambda) = \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K X_{ikl} \right|^2,$$

$$\hat{f}_{x_{ik} - \bar{x}_i, x_{ik} - \bar{x}_i}(\lambda) = \frac{1}{L} \sum_{l=1}^L \left| X_{ikl} - \frac{1}{K} \sum_{k=1}^K X_{ikl} \right|^2$$

and

$$\hat{f}_{\bar{x}_i - \bar{x}., \bar{x}_i - \bar{x}..}(\lambda) = \frac{1}{L} \sum_{l=1}^L \left| \frac{1}{K} \sum_{k=1}^K X_{ikl} - \frac{1}{IK} \sum_{i=1}^I \sum_{k=1}^K X_{ikl} \right|^2.$$

This method assumes a constant series length, T and a constant number of replications, K . To satisfy these constraints the data obtained from Person 8 is removed, as only 1 time series is available for each product and 6 time series for each of the remaining (i, j) combinations are taken, resulting in a loss of 8 time series with a new total of 216. Since it is not possible to extend any of the time series, the constant T must be taken as the minimum of the remaining 216 series lengths, namely $T = 112$. Since the true starting and finishing points of the application process, m_2 and m_3 , may not have been correctly estimated it was decided that each time series longer than 112 observations would have an equal number of observations removed from the beginning and end of the series such that only the central 112 observations were retained for analysis. It was also noted that if Person 10 was also removed, the minimum series length of the remaining time series would almost double, giving $T = 212$. Thus the analysis was repeated without the 24 time series from Person 10, giving a new total of 192 time series, each of length 212 observations, retained for

study. Again, the 212 observations were taken from the central portion of each time series. A choice of $L = 10$, equal to the value given in Brillinger [1972], is studied first and then the analysis is repeated with $L = 5$ to see if this will alter the results.

The model given in equation 7.8 shall be referred to as Model 1. The analysis was then repeated for the model

$$x_{ijk}(t) = \mu_{ijk} + a(t) + c_j(t) + \varepsilon_{ijk}(t)$$

where μ_{ijk} are constants, $a(t)$ is a stationary time series common to all x_{ijk} , the $c_j(t)$ are stationary time series common to each product and the $\varepsilon_{ijk}(t)$ are stationary time series unique to each x_{ijk} . This exchanges the assumption of no product effect with the assumption instead that no person effect is present. This model shall be referred to as Model 2. The details for Model 2 are similar to those for Model 1 and thus omitted.

Brillinger [1972] also mentions that this method can be extended to include random effects data collected in more complicated experimental designs in a straight forward manner. It is thus extended to the combined effect model of

$$x_{ijk}(t) = \mu_{ijk} + a(t) + b_i(t) + c_j(t) + d_{ij}(t) + \varepsilon_{ijk}(t)$$

where μ_{ijk} are constants, $a(t)$ is a stationary time series common to all x_{ijk} , the $b_i(t)$ are stationary time series common to each person, the $c_j(t)$ are stationary time series common to each product, the $d_{ij}(t)$ are stationary time series common to each (i, j) combination and the $\varepsilon_{ijk}(t)$ are stationary time series unique to each x_{ijk} . This shall be referred to as Model 3.

Using the same arguments as before, four F statistics can be produced for each value of λ for Model 3. The first, under the null hypothesis that $f_{aa}(\lambda) = 0$, is given by

$$\frac{IJK \hat{f}_{\bar{x} \dots, \bar{x} \dots}(\lambda)}{\sum_i \sum_j \sum_k \hat{f}_{x_{ijk} - \bar{x}_{ij}, x_{ijk} - \bar{x}_{ij}}(\lambda) / (IJ(K-1))} \sim F_{2L, 2IJ(K-1)L}^1$$

The second, under the null hypothesis that $f_{bb}(\lambda) = 0$, is given by

$$\frac{JK \sum_i \hat{f}_{\bar{x}_{i \dots}, -\bar{x} \dots, \bar{x}_{i \dots}, -\bar{x} \dots}(\lambda) / (I-1)}{\sum_i \sum_j \sum_k \hat{f}_{x_{ijk} - \bar{x}_{ij}, x_{ijk} - \bar{x}_{ij}}(\lambda) / (IJ(K-1))} \sim F_{2(I-1)L, 2IJ(K-1)L}^1$$

The third, under the null hypothesis that $f_{cc}(\lambda) = 0$, is given by

$$\frac{IK \sum_j \hat{f}_{\bar{x}_{i \dots}, -\bar{x} \dots, \bar{x}_{i \dots}, -\bar{x} \dots}(\lambda) / (J-1)}{\sum_i \sum_j \sum_k \hat{f}_{x_{ijk} - \bar{x}_{ij}, x_{ijk} - \bar{x}_{ij}}(\lambda) / (IJ(K-1))} \sim F_{2(J-1)L, 2IJ(K-1)L}^1$$

The fourth, under the null hypothesis that $f_{dd}(\lambda) = 0$, is given by

$$\frac{K \sum_i \sum_j \hat{f}_{\bar{x}_{ij} - \bar{x}_{i..} - \bar{x}_{.j} + \bar{x}_{...}, \bar{x}_{ij} - \bar{x}_{i..} - \bar{x}_{.j} + \bar{x}_{...}}(\lambda) / ((I-1)(J-1))}{\sum_i \sum_j \sum_k \hat{f}_{x_{ijk} - \bar{x}_{ij}, x_{ijk} - \bar{x}_{ij}}(\lambda) / (IJ(K-1))} \sim F_{2(I-1)(J-1)L, 2IJ(K-1)L}$$

For each, if the F statistic is larger than the critical F value for a given significance level the null hypothesis is rejected and significant evidence is present to suggest the corresponding series of $a(t)$, $b_i(t)$, $c_j(t)$ or d_{ij} is significant. Once again this analysis is carried out without the data from Person 8 and with 6 time series for each of the remaining (i, j) combinations, resulting in $I = 9, J = 4, K = 6, T = 112$. The analysis is then repeated without Person 10 such that $I = 8, J = 4, K = 6, T = 212$. For each the choices of $L = 10$ and $L = 5$ are made and the results between the two compared.

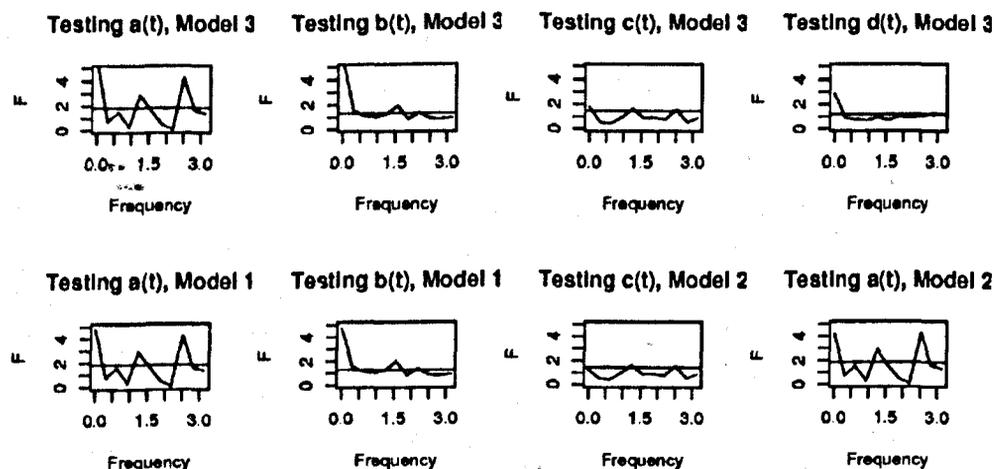


Fig. 7.4: The F statistics for each frequency $\lambda \in (0, \pi)$ for the 216 series from 9 individuals with $T = 112$ and $L = 5$. The horizontal lines represent the critical F values at the 95% significance level.

The F statistics and corresponding critical F values at the 95% significance level for each frequency $\lambda \in (0, \pi)$ for the 216 series from 9 individuals with $T = 112$ and $L = 5$ are given in figure 7.4. It can be seen that the F statistics do not significantly change depending on which model is used. This is expected due to the independence of the F statistics in the combined model. This similarity is also seen for the other values of L and T and thus the remaining results for the Models 1 and 2 are omitted.

From figure 7.4, there appears to be significant evidence to suggest the effect of the individual is present at frequencies near $\lambda = 0$ and that an interaction effect between individual and product may also be present at low frequencies. No strong evidence of effects of individual or product seem present away from $\lambda = 0$, other than

a smaller peak in the F statistic near $\lambda = 1.5$ for the person effect. The component $a(t)$ common to all the series also appears significant at low frequencies and appears more significant than the individual effect near $\lambda = 1.5$ and $\lambda = 2.6$.

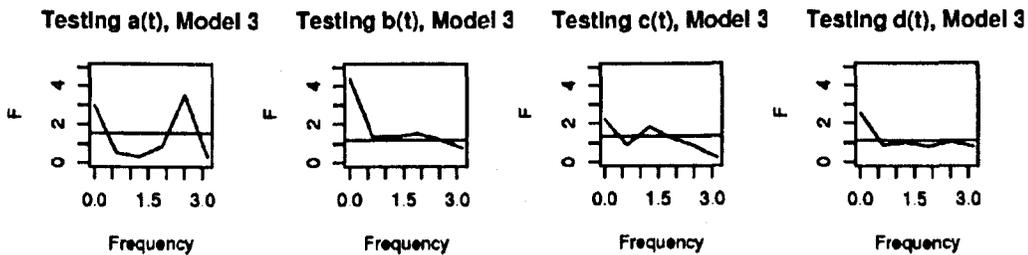


Fig. 7.5: The F statistics for each frequency $\lambda \in (0, \pi)$ for the 216 series from 9 individuals with $T = 112$ and $L = 10$. The horizontal lines represent the critical F values at the 95% significance level.

Figure 7.5 shows similar results near the low frequencies are found when L is taken as 10 instead of 5 in that it seems to show significant evidence of the presence of an individual effect, an interaction between individual and product and a component common to all. In addition, the product effect also shows some significance at the 95% level for this value of L .

The significant peak near $\lambda = 2.6$ for the common component is still present for this value of L , but there is no longer any significant evidence to suggest the presence of a common frequency near $\lambda = 1.5$. For this value of L the F statistic for the individual effect is generally just above the critical values at the 95% level and thus significant, whereas for $L = 5$ the F statistic for the individual effect is generally just below. This shows a sensitivity of the results to the choice of L .

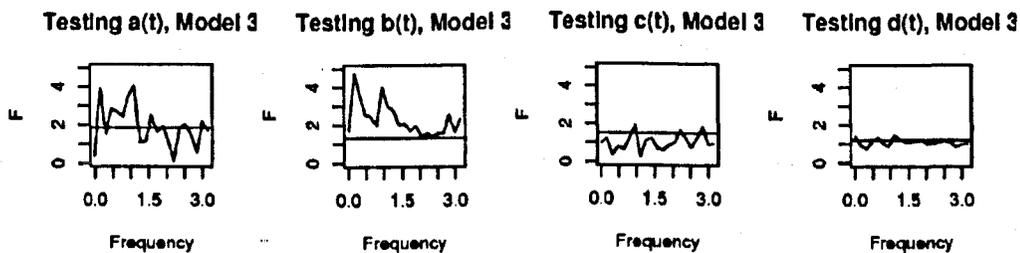


Fig. 7.6: The F statistics for each frequency $\lambda \in (0, \pi)$ for the 192 series from 8 individuals with $T = 212$ and $L = 5$. The horizontal lines represent the critical F values at the 95% significance level.

With the removal of person 10, the value of T can be increased from 112 to 212. Figure 7.6 gives the F statistics and corresponding critical values at the 95%

significance level for the remaining 8 individuals with the increased series length, $T' = 212$, and $L = 5$. Comparison with figure 7.4 shows the effect of the individual now appears far more significant, being above the 95% significance level critical value for every value of λ . This suggests a series length of $T' = 112$ may have been too small to pick up on this. The component $a(t)$ common to all series again shows signs of significance, but the frequencies at which this significance occurs are different from those found for $T' = 112$. The product and interaction effects show no significant evidence of being present for this series length even at low frequencies.

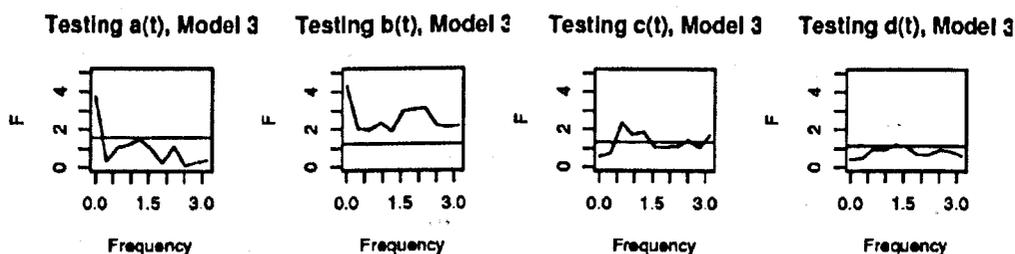


Fig. 7.7: The F statistics for each frequency $\lambda \in (0, \pi)$ for the 192 series from 8 individuals with $T' = 212$ and $L = 10$. The horizontal lines represent the critical F values at the 95% significance level.

Figure 7.7 shows the F statistics for the remaining 8 individuals with $T' = 212$ and $L = 10$. As with the smaller series length of $T' = 112$, the results still appear to show some sensitivity to the choice of L . Comparison of figure 7.7 with figure 7.6 shows that although the effect of the individual is significant for all λ in both, the choice of $L = 10$ seems to increase this significance at higher frequencies. The common component shows less significance for this value of L , with significant evidence to suggest its presence only at small frequencies. The product and interaction effects are generally not significant, although this choice of L gives some signs of significance for the product effect near $\lambda = 1$.

Overall, the presence of a common component to all series seems significant near low frequencies, though its significance away from $\lambda = 0$ is questionable due to contradictory results dependent on L and T' . The presence of an individual effect seems significant at least at low frequencies and the results for $T' = 212$ suggest significance for the majority of frequencies regardless of the choice of L . The results for the product and interaction effects show some small signs of significance, but these once again seem largely dependent on the choice of L and no consistent signs of significance can be seen.

The results here are thus in agreement with those of the time domain ANOVA results given in section 7.2.2, suggesting evidence of different people having different application techniques but suggesting that differences between products are not clear from this data.

7.3 Multivariate Analysis

During this section, the focus of the analysis is returned to the entire data sets Y_{ijk} defined in section 7.1. The notation $\{a, b\}_{ijk}$ will again be used to represent the time series generated by the distances between sensors a and b for the ijk th trial, where the sensors are numbered by the list given in section 7.1 and the subtext ijk will be dropped when not necessary.

Since there are 224 sets of data, each containing 21 time series, this gives a total of 4704 time series of varying lengths to look at. The use of data reduction tools are thus desirable and section 7.3.1 looks at the use of Principal Component Analysis (PCA) to achieve this. Section 7.3.3 then attempts to reconstruct meaningful (x, y, z) coordinates such that the data can be more easily interpreted and comparisons between trials can be made.

7.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a data-analytic technique used to describe a multivariate structure. PCA was originally considered by Pearson [1901] in relation to two variables, fitting a line to a scatter plot for the case in which x and y were co-dependent. This line is now known as the line of best fit. This method was extended to the method of principal components for multivariate data, where the line of best fit between two variables is the first principal component (PC), see Hotelling [1933].

The use of PCA has been applied to many varied areas of research, including examples of psychology and education found in Harmon [1976], quality control found in Fisher et al. [1986], chemistry found in Weiner [1973], photography found in Simonds [1963], market research found in Vavra [1972] and economics found in Bartlett [1948]. Several other examples of areas of application for PCA can be found in Jackson [1991].

The method of PCA relies on the following result from matrix algebra.

Theorem 7.3.1. *Let S be a $p \times p$ symmetric, non-singular matrix. Then there exists an orthonormal matrix U such that*

$$U'SU = L$$

where

$$U'U = I_p$$

and L is a diagonal matrix with elements l_1, l_2, \dots, l_p . Also

$$|S| = |L| = l_1 l_2 \dots l_p$$

and

$$\text{Tr}S = \text{Tr}L = l_1 + l_2 + \cdots + l_p.$$

The columns of U , u_1, u_2, \dots, u_p , are known as the *characteristic vectors* or *eigenvectors* of S . The diagonal elements of L , l_1, l_2, \dots, l_p , are known as the *characteristic roots* or *eigenvalues*.

Now, let S be the covariance matrix of p variables $\mathbf{x} = (x_1, \dots, x_p)'$ and let U be the matrix of eigenvectors of S . The PCs of \mathbf{x} are given by \mathbf{z} , where

$$\mathbf{z} = U'(\mathbf{x} - \bar{\mathbf{x}}),$$

and $\bar{\mathbf{x}}$ is the vector of the means of \mathbf{x} . These PCs are unique, and given the vector \mathbf{z} , the original variables can be found via

$$\mathbf{x} = U\mathbf{z} + \bar{\mathbf{x}},$$

from the orthonormal property of U . Note, the covariance matrix, S_z , of \mathbf{z} is given by

$$S_z = U'SU = L.$$

Hence, the PCs are an uncorrelated transformation of the original variables, with $\text{Var}(z_i) = l_i$ for $i \in (1, \dots, p)$. From Theorem 7.3.1 it can be seen that

$$\sum_{i=1}^p \text{Var}(x_i) = \sum_{i=1}^p \text{Var}(z_i).$$

The PCs are thus ordered such that $l_1 > l_2 > \cdots > l_p$ so that the first PC accounts for the largest proportion of the total variance of the original variables, and the last PC accounts for the smallest.

If several of the original variables are strongly correlated, the majority of the variance of these variables may be able to be explained by relatively few PCs, providing a reduction in dimensionality and possibly aiding in interpreting the main sources of the variability. The use of PCA to reduce the dimensionality writes \mathbf{x} in the form,

$$\mathbf{x} = \bar{\mathbf{x}} + U_k \mathbf{z}_k + \Sigma,$$

where U_k and \mathbf{z}_k are the first k eigenvectors and PCs for $k < p$ and Σ is a vector of the error terms $\varepsilon_1, \dots, \varepsilon_p$, each with zero mean. If the vector Σ can be assumed negligible, then the dimensionality has been reduced from p to k .

Choosing the number of PCs to retain

Various methods of choosing k have been developed, and a brief introduction to some of these methods is presented here. For more thorough reviews of these techniques the reader is referred to Jackson [1991] section 2.8 and Jolliffe [2002] chapter 6.

The first method presented here is to choose the smallest k such that a certain proportion, C , of the original variances is explained by the retained PCs. This method is often used due to its simplicity. However, since the choice of what proportion of the variances to account for is somewhat arbitrary, this method may not result in an ideal selection of k . Jolliffe [2002] suggests smaller values of C may be appropriate when p is large, whilst larger values of C may be required if the first few PCs dominant but smaller, less obvious structures are still of interest.

A related method of choosing k is to instead choose the smallest k such that the variances of the error terms are below a certain threshold, C . This can either be done for each error term separately, such that

$$\text{Var}(\varepsilon_i) \leq C, \quad \text{for all } i \in (1, \dots, p), \quad (7.12)$$

or for the sum of the error variances,

$$\sum_{i=1}^p \text{Var}(\varepsilon_i) \leq pC. \quad (7.13)$$

Note, equation 7.12 implies equation 7.13 is true, thus the choice of k that satisfies the second will be less than or equal that of the first.

The next method looks at the individual variances of the PCs. It suggests removal of all PCs with a variance less than the average, \bar{l} . This method is sometimes referred to as the *Guttman-Kaiser* criterion, having been developed by Guttman [1954] and Kaiser [1960]. Simulation studies by Jolliffe [1972] showed this method tended to retain too few PCs, and suggested instead the use of $l^* = 0.7\bar{l}$ as the cut-off for the variances. Jolliffe [1986] extended this idea to develop the *broken stick* method. The idea is that if a stick of unit length were randomly broken into p segments, the expected length of the k th-longest segment would be

$$g_k = \frac{1}{p} \sum_{i=k}^p \frac{1}{i}.$$

Hence, the expected proportion of the variance explained by the k th largest PC would also be g_k by chance alone. The broken stick method thus retains only those PCs for which the proportion of the variance explained is greater than the corresponding value of g_k .

Graphical methods of determining the number of PCs to retain are also often used. A plot of l_i against i , called a SCREE plot by Cattell [1966], tends to show

that l_i for $i > k$ tend to be small and lie close to a straight line. The method of choosing k is to determine by inspection the point at which this break occurs. Cattell and Jaspers [1966] suggest retaining all PCs up to and including the first one after this break.

There are several difficulties with this method. The location of the break may not be obvious, or indeed, no break may be present and the plot could appear like a fairly smooth curve. Another difficulty may arise when there appears to be more than one break. Jackson [1991] suggests it is customary to use the first break in this case, but gives examples in which later breaks may be of interest. Also, if the first few values of l_i are much larger, it may be difficult from a plot to see the detail required to determine the break point. This difficulty can be reduced by a plot instead of $\log(l_i)$ against i . These plots are known as *log-eigenvalue* (LEV) plots. A study of these plots can be found in Farmer [1971].

Finally, several statistical tests have been developed to estimate k , see for example Bartlett [1950] and Bentler and Yuan [1996],[1998]. However, in order to use these tests to find a suitable k , they must generally be repeatedly applied to remove or include PCs one at a time. Since multiple tests are performed, this implies the overall significance level will not be equal to that of each individual test. Further, since the number of tests to be carried out is random, it makes the true overall significance level of the tests difficult to find. For the test of Bartlett [1950] in particular, Jolliffe [1970] found that this method tended to retain too many PCs, performing similar to using the Guttman-Kaiser criterion with a cut-off of around $l^* = 0.1\bar{l}$ or $0.2\bar{l}$, far smaller cut-off than the optimal value suggested in Jolliffe [1972] of $l^* = 0.7\bar{l}$. Jackson [1991] thus suggests use of this test to find a maximum value of k .

7.3.2 PCA of Y_{ijk}

Let Y_{ijk} be the form of the deodorant stick data that represents the distances between sensors, defined in section 7.1. PCA is now applied to Y_{ijk} in an attempt to reduce the dimensionality.

As described in section 7.3.1, PCA writes the multivariate time series Y_{ijk} in the form

$$Y_{ijk}(t) = R_{ijk}P_{ijk}(t) + \Sigma_{ijk}(t),$$

where $Y_{ijk}(t)$ is a 21×1 vector, R_{ijk} is a $21 \times r$ matrix of coefficients, $P_{ijk}(t)$ is a $r \times 1$ vector of PCs and $\Sigma_{ijk}(t)$ is a 21×1 vector of residuals, uncorrelated with $P_{ijk}(t)$. The value r is the number of PCs to be retained such that $0 \leq r \leq 21$. The notation r has been used instead of the k used in section 7.3.1 to avoid confusion with the repetition number.

Let $Y(t), t \in (\hat{m}_2, \dots, \hat{m}_3)$ be the first set of multivariate data such that $i, j, k = 1$

and $\hat{m}_2 = 146$, $\hat{m}_3 = 1323$ are the estimated start and finish points of the deodorant stick application found using the method described in section 7.2.1.

The 1st PC of $Y(t)$ accounts for 72% of the total variance of $Y(t)$, the 2nd PC accounts for 14% and the 3rd PC accounts for a further 6%. In order to account for over 70% of the total variance r is thus chosen as 1, whilst to account for over 80% r is taken as 2 and to account for over 90% r is taken as 3.

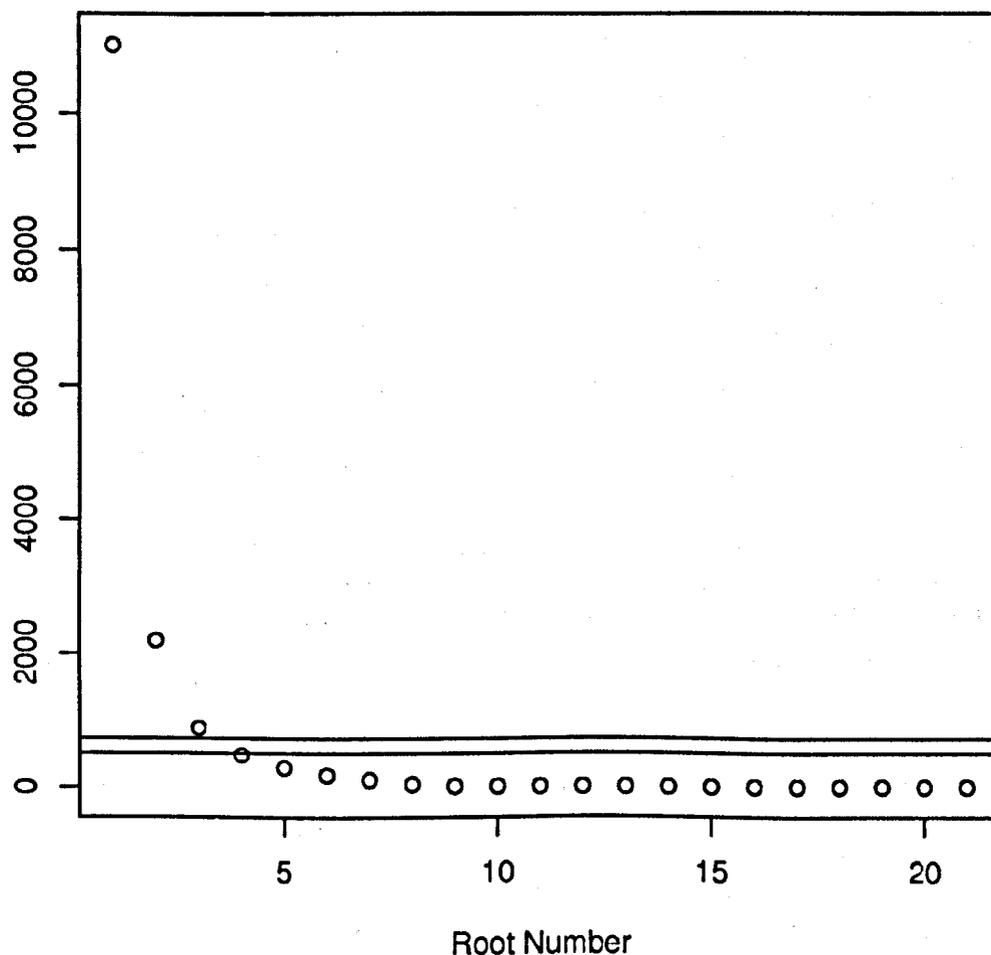


Fig. 7.8: The SCREE plot for $Y(t), t \in (\hat{m}_3, \dots, \hat{m}_4)$. The horizontal lines represent the mean of the characteristic roots, \bar{l} and $l^* = 0.7\bar{l}$.

The SCREE plot for the characteristics roots are given in figure 7.8. There appears to be possibly two breaks, the first near $i = 4$ and the latter near $i = 8$, although the location of these are not obvious. This suggests perhaps retaining $r = 4$ PCs. Figure 7.8 also shows the values of \bar{l} and $l^* = 0.7\bar{l}$ used for the Guttman-

Kaiser criterion, which both suggest taking $r = 3$. The broken stick method suggests taking $r = 2$, with only the first two eigenvalues explaining greater than the expected proportion of the overall variance.

These methods suggest retaining at most 4 of the PCs, showing a large reduction in the overall number of series required.

As each of the 21 series present in $Y(t)$ is the distance between a pair of moving sensors, interpretation of these two PCs becomes difficult. In an attempt to aide with the interpretation, PCs of various subsets of $Y(t)$ were also taken and compared with the two PCs of the full set to find simpler to interpret approximations. Although initial choices of which subsets to study could be made from the matrix of rotation coefficients, R , by excluding series with coefficients close to zero, some use of trial and error was also required. It was assumed that the first PC may be related to the movement of sensor 1, as this is at the head of the moving deodorant stick. Looking therefore at subsets involving distances from sensor 1 it was found that the first PC, $P_1(t)$, could be approximated by

$$\tilde{P}_1(t) = -0.4\{1, 3\} - 0.5\{1, 4\} + 0.5\{1, 5\} + 0.5\{1, 6\} - 0.4\{1, 7\}. \quad (7.14)$$

Looking similarly at subsets involving each of the other sensors, it was found that the 2nd PC, $P_2(t)$, could be approximated by

$$\tilde{P}_2(t) = 0.6\{3, 5\} + 0.8\{3, 6\} + 0.1\{3, 7\}. \quad (7.15)$$

Figure 7.9 shows how close the approximations appear to the actual PCs of $Y(t)$. The correlation between $\tilde{P}_1(t)$ and $P_1(t)$ was 0.995, whilst the correlation between $\tilde{P}_2(t)$ and $P_2(t)$ was 0.826. Looking at equation 7.14, it is apparent that that each of the series included in $\tilde{P}_1(t)$ involve distances from sensor 1 and hence the first PC may indeed be related to the movement of sensor 1, which is used to represent the movement of the deodorant stick. Further interpretation as to how $\tilde{P}_1(t)$ may measure this movement is somewhat vague, but since it is weighted sum of the distances from sensor 1 to sensors 5 and 6, which can generally be considered to be 'below' sensor 1, minus the weighted sum of the distances from sensor 1 to sensors 3, 4 and 6, which can generally be considered to be 'above' sensor 1, it suggests $\tilde{P}_1(t)$ represents a form of 'up and down' movement of sensor 1. This movement, however, can not be assumed to be in any fixed direction relevant to the individual.

Looking similarly at $\tilde{P}_2(t)$, it appears that the 2nd PC is related to the movement of sensor 3, the sensor attached to the individual's wrist. It is a weighted sum of the distances from sensor 3 to sensors 5, 6 and 7, the three sensors attached to the individual's side and shoulder. The 2nd PC thus seems to relate to the raising and lowering of the wrist away from and towards the individual. Again, the interpretation is vague and no fixed direction can be assumed.

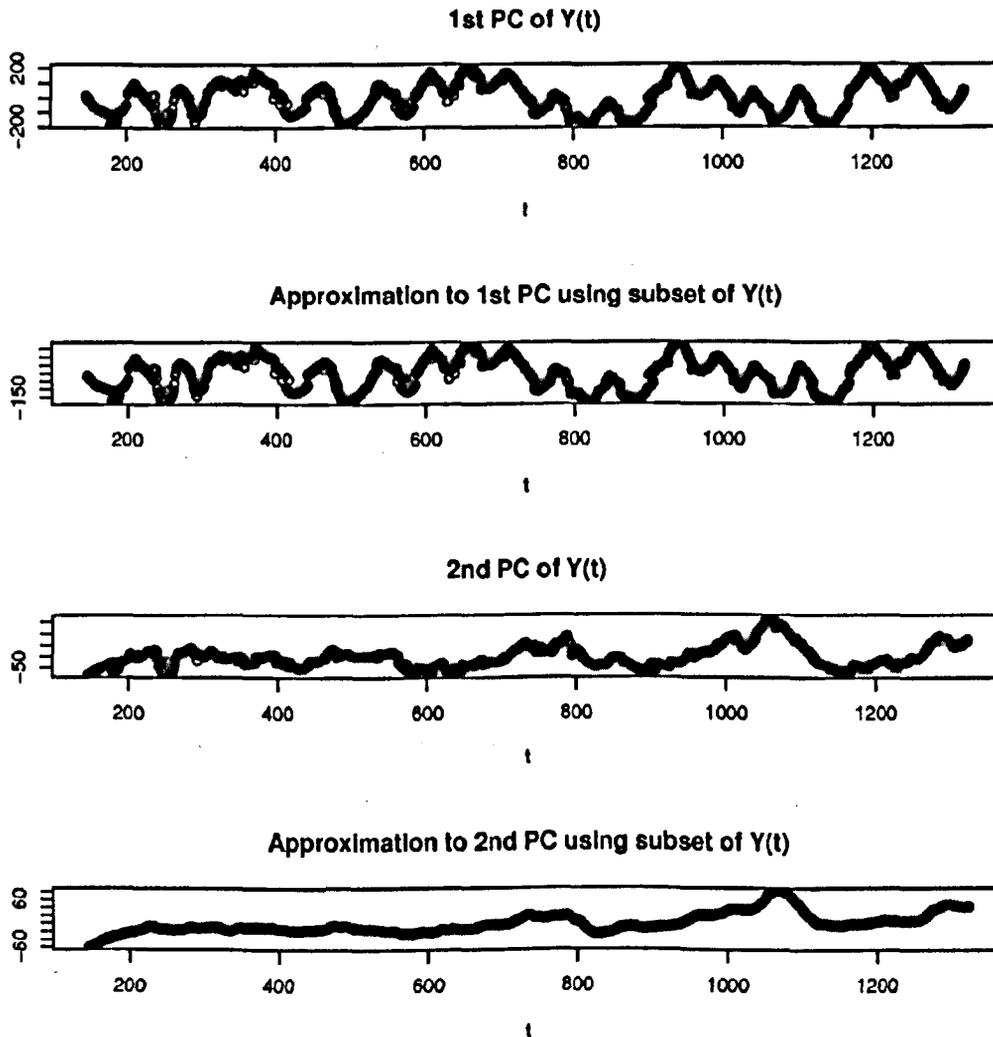


Fig. 7.9: The 1st and 2nd PCs of $Y(t)$ and the corresponding approximations given in equations 7.14 and 7.15 respectively.

Close approximations to the first two PCs of $Y(t)$ can also be made from series involving sensors 2 and 4 respectively instead of sensors 1 and 3, although these are inferior, with correlations of 0.779 and 0.813 respectively. This seems to be because the two sensors on the deodorant stick and the two sensors on the individual's arm appear to move in a similar fashion for this trial and are therefore more or less interchangeable.

The PCA was repeated for all 224 trials and the proportion of the variance explained by each PC was recorded for all. The minimum, maximum, mean and standard deviations of the proportions of the variances explained by the first four PCs can be seen in table 7.3. The table shows that to explain over 90% of the original variance, the number of PCs retained should be taken as $r \in (1, 2, 3, 4)$, depending on the trial, with a value of $r = 2$ being sufficient on average.

	Separately				Cumulatively			
	Min	Max	Mean	SD	Min	Max	Mean	SD
1st PC	0.446	0.948	0.735	0.128	0.446	0.948	0.735	0.128
2nd PC	0.025	0.410	0.175	0.095	0.721	0.976	0.910	0.052
3rd PC	0.013	0.201	0.051	0.032	0.866	0.992	0.961	0.026
4th PC	0.003	0.072	0.019	0.013	0.926	0.996	0.981	0.014

Tab. 7.3: Proportion of variance explained by PCs for Y_{ijk} .

Table 7.4 shows the number of trials each value of r was chosen for several criteria mentioned in section 7.3.1. The maximum number of PCs that any of these methods suggest to retain is $r = 5$, although this occurred for only 1 trial using the Guttman-Kaiser criterion with the cut-off at 0.7 \bar{l} . The modal value of the estimated r using the cut-off points of 70% of the total variances is 1, the modal value using the Guttman-Kaiser criterion with the cut-off at 0.7 \bar{l} is $r = 3$ and the modal values for the other methods are all $r = 2$. The mean values of r for each method are between 1 and 3. These results suggest a large reduction in dimensionality can be obtained for all trials.

Having determined that each of the trials generally appears to require $r \leq 4$ trials, two further individual trials are looked at to discover if there are similarities in the interpretation of these retained PCs between trials. Let $Y_2(t)$, $t \in (\hat{m}_{22}, \dots, \hat{m}_{32})$ be the set of data for which the first four PCs explain the smallest proportion of the overall variance of all the trials. This is the trial corresponding to $i = 9, j = 1, k = 2$, with $\hat{m}_{22} = 142$ and $\hat{m}_{32} = 507$. Also, let $Y_3(t)$, $t \in (\hat{m}_{23}, \dots, \hat{m}_{33})$ be the set of data for which the first two PCs explain the largest proportion of the overall variance of all the trials. This is the trial corresponding to $i = 5, j = 1, k = 5$, with $\hat{m}_{33} = 163$ and $\hat{m}_{43} = 412$. As seen in table 7.3, the first 4 PCs of $Y_2(t)$ account for 93% of the total variance for this set of data, whilst the first 2 PCs of $Y_3(t)$ account for 98% of the total variance for this set. The values of r chosen using the cut-off points of

\hat{r}	Cut-Off Point			Guttman-Kaiser		Broken Stick
	70%	80%	90%	l	$0.7l$	
1	130	90	20	10	2	85
2	94	122	131	113	83	123
3	0	12	67	92	101	16
4	0	0	6	9	37	0
5	0	0	0	0	1	0
mean	1.42	1.65	2.26	2.45	2.79	1.69
SD	0.49	0.58	0.65	0.65	0.74	0.60

Tab. 7.4: The number of trials each value of r was chosen using the cut-off points of 70%, 80% and 90% of the total variances, the Guttman-Kaiser criterion using l and $0.7l$ and the broken stick method. The rows labelled mean and SD give the mean and standard deviations of the chosen r for each method.

70%, 80% and 90% of the total variances, the Guttman-Kaiser criterion using l and $0.7l$ and the broken stick method for $Y_2(t)$ were $r = 2, 3, 4, 3, 3$ and 2 and for $Y_3(t)$ were $r = 1, 1, 1, 2, 2$ and 1 respectively.

As with the PCs for $Y(t)$, approximations of the PCs for $Y_2(t)$ and $Y_3(t)$ were made using subsets of each. The 4 PCs of $Y_2(t)$ were approximated by

$$\tilde{P}_{12}(t) = -0.4\{1, 3\} - 0.5\{1, 4\} + 0.5\{1, 5\} + 0.5\{1, 6\} - 0.3\{1, 7\},$$

$$\tilde{P}_{22}(t) = 0.9\{2, 3\} + 0.4\{2, 4\} - 0.1\{2, 6\},$$

$$\tilde{P}_{32}(t) = -0.5\{2, 4\} + 0.5\{2, 5\} + 0.7\{2, 6\}$$

and

$$\tilde{P}_{42}(t) = -0.6\{1, 4\} + 0.8\{1, 7\}.$$

The correlations of these series with the original four PCs are 0.999, 0.948, 0.949 and 0.897 respectively. The form of $\tilde{P}_{12}(t)$ can be seen to be very similar to that of $\tilde{P}_1(t)$ given in equation 7.14, thus the 1st PC seems to once again give a vague measure of the 'up and down' motion of sensor 1. The approximations to the next two PCs involve only distances from sensor 2. This suggests the next two PCs are related to the movement of sensor 2, though perhaps in two different general directions. The form of $\tilde{P}_{22}(t)$ suggests the 2nd PC may be the movement of sensor 2 away from the arm, whilst the form of $\tilde{P}_{32}(t)$ suggests the 3rd PC may be the movement of sensor 2 away from the individual's side. The form of $\tilde{P}_{42}(t)$ again involves distances from sensor 1, suggesting the 4th PC may be the movement of sensor 1 along the direction connecting sensors 4 and 7, the sensors on the elbow and

shoulder. Note, this direction is still not fixed as both sensors are freely moving, but if the arm is sticking out to the side it may relate more to a horizontal movement of sensor 1, as opposed to the vertical movement suggested by the 1st PC.

The 2 PCs of $Y_3(t)$ were approximated by

$$\tilde{P}_{13}(t) = -0.5\{1,3\} - 0.5\{1,4\} + 0.4\{1,5\} + 0.5\{1,6\} - 0.2\{1,7\}$$

and

$$\tilde{P}_{23}(t) = 0.4\{1,6\} + 0.6\{2,3\} + 0.7\{3,6\}. \quad (7.16)$$

The correlations of these series with the original two PCs are 0.995 and 0.989 respectively. The form of $\tilde{P}_{13}(t)$ can also be seen to be very similar to that of $\tilde{P}_1(t)$, suggesting the 'up and down' motion of sensor 1 is once again mainly responsible for the 1st PC of $Y_3(t)$. A reasonable approximation in appearance to the 2nd PC of $Y_3(t)$ was not found in terms of distances from one sensor. The form of $\tilde{P}_{23}(t)$ is difficult to meaningfully interpret even in a vague sense.

In general, the 1st PC for each $Y_{ijk}(t) \ t \in (\hat{m}_{2ijk}), \dots, \hat{m}_{3ijk}$, $i \in (1, \dots, 10)$, $j \in (1, \dots, 4)$, $k \in (1, \dots, K_{ij})$ appeared very similar to the first PC of the subsets of the series $\{1,3\}, \{1,4\}, \{1,5\}, \{1,6\}$ and $\{1,7\}$. The minimum correlation between the two series for any trial was 0.613, although only 12 trials had a correlation of less than 0.8. In comparison, 189 of the trials had a correlation of greater than 0.95 and 116 of the trials had a correlation of greater than 0.99, with a maximum correlation of 0.999972. The form of P given below in equation 7.17, gives the average form of these approximations to the 1st PCs, with the coefficients of P being the mean of the coefficients over the 224 trials.

$$P = \begin{matrix} -0.45\{1,3\} & - & 0.47\{1,4\} & + & 0.47\{1,5\} & + & 0.48\{1,6\} & - & 0.34\{1,7\}. \\ (0.03) & & (0.03) & & (0.02) & & (0.02) & & (0.1) \end{matrix} \quad (7.17)$$

The values in brackets are the standard deviations of the estimated coefficients over the 224 trials. It can be seen that the variances of the first four coefficients are all small, whilst the variance of the last coefficient is relatively large in comparison. It is believed this is due to sensor 7 not always being 'above' sensor 1 as the individuals may on occasion raise the deodorant stick above the shoulder level.

The results of the PCA of the data sets $Y_{ijk}(t)$ have shown that a large reduction in dimensionality can be achieved whilst still accounting for over 90% of the variances. The 1st PC for each, accounting for between 45% and 95 % of the total variance, appears to be related to the movement of sensor 1, though the direction of this movement is vague. The remaining PCs often also appear to be associated with the movement of a particular sensor or pair of sensors, though again no fixed directions for these movements can be claimed. For some of the trials, not even a vague description of the PCs can be meaningfully found.

The difficulty arises in the form of the data sets $Y_{ijk}(t)$, as even the original data is difficult to interpret in terms of how each sensor is moving. Section 7.3.3 attempts to overcome this difficulty by transforming the data into a more suitable form.

7.3.3 Graphical Representation

The original raw data, X_{ijk} , defined in section 7.1, contained the (x, y, z) coordinates of each of the seven sensors, however the axes from which these coordinates were measured did not remain fixed between trials, making comparisons difficult. The data was thus transformed into Y_{ijk} , which consists of vectors $Y_{ijk}(t)$ for $t \in (1, \dots, T_{ijk})$ such that each $Y_{ijk}(t)$ is a vector of the 21 between sensor distances at the time t .

The PCA on these $Y_{ijk}(t)$, carried out in section 7.3.2, showed that the interpretation of the data in this form, however, can often be difficult. This section overcomes the difficulties of both X_{ijk} and Y_{ijk} by transforming the data back into a form of coordinates by creating axes from the data which are fixed relative to the sensors.

A set of axes for a 3 dimensional space can be defined by three points in space. The first point to be defined is the origin, $(0,0,0)$. The simplest method to define this point is to take one of the sensors and fix it at this location. Out of the seven sensors, the sensor most likely to remain in a constant position is sensor 6, the lowest sensor on the individuals side, since sensors 1 and 2 are on the moving deodorant stick, sensors 3 and 4 are on a possibly moving arm and sensors 5 and 7 are closer to the moving parts, thus more likely to be effected by them. Thus, for all trials at all times, sensor 6 is set to the origin $(0,0,0)$ and all other positions are measured from this one.

The second point should be a point in the direction of an axis. Again, since sensors 1,2,3 and 4 may all be moving considerably, only sensors 5 and 7 need be considered. Since sensor 5 is closer to sensor 6 small variations would have a greater effect on the direction of the line. Therefore sensor 7 is chosen to define the 'up' or y axis. This is done by setting sensor 7 to the position $(0, \{6,7\}_t, 0)$ for all trials at all times. Note the distance between sensors 6 and 7 is still allowed to vary and that the position of sensor 7 has not been fixed, only the y axis has been fixed relative to this position.

Finally, the third point should finish the definition of the x, y plane and the z axis can be taken as the normal to this plane without further points required. Sensor 5 is the most straight forward choice for this. In addition to the previous reasons which make sensors 1,2,3 and 4 illogical, sensor 5 gives an easier interpretation of the x, y plane. Since sensors 5, 6 and 7 all lie along the individuals side, the x, y plane can be taken to be a representation of this individuals side. Since it would be desirable for the x and y axes to be at right angles to each other, the position of sensor 5 is not set to $(\{5,6\}_t, 0, 0)$, but instead set to $(a_t, b_t, 0)$, where a_t is the distance of sensor 5 from the newly defined y axis and b_t is its projection onto this

y axis. Using standard techniques, these are given by

$$b_t = \frac{\{5,6\}_t^2 + \{6,7\}_t^2 - \{5,7\}_t^2}{2\{6,7\}_t} \quad (7.18)$$

and

$$a_t = \sqrt{\{5,6\}_t^2 - b_t^2}. \quad (7.19)$$

The z axis is then taken as the normal to this plane at the origin such that the positive z direction is away from the individual. Note that these new axes are not fixed in space. Instead, they are free to move with the individual, with the x, y plane always along the individuals side, regardless of how the individual may stand or move during application.

For any of the remaining sensors, the location of sensor s on the new axes can be found using the following formulae based on standard geometric techniques.

Let s be located at the point (x_t, y_t, z_t) at time t . Then

$$y_t = \frac{\{s,6\}_t^2 + \{6,7\}_t^2 - \{s,7\}_t^2}{2\{6,7\}_t},$$

is the projection of s on the y axis. Now, let c_t be the distance from the projection of s onto the line joining sensors 5 and 6 to the origin,

$$c_t = \frac{\{s,6\}_t^2 + \{5,6\}_t^2 - \{s,5\}_t^2}{2\{5,6\}_t},$$

and a_t and b_t be defined as in equation 7.19 and equation 7.18 respectively. Then

$$x_t = \frac{c_t \sqrt{a_t^2 + b_t^2} - b_t y_t}{a_t}$$

and

$$z_t = \sqrt{\{s,6\}_t^2 - x_t^2 - y_t^2}.$$

The data can now be represented once again in terms of (x, y, z) coordinates of each of the sensors. Notice that the three time series describing the location of sensor 6 will all be zero constants and can thus be disregarded. Notice also that the same is true for the x and z series of sensor 7 and the z series of sensor 5 and these can also be disregarded, hence the total number of time series for each trial is 15 rather than 21. The notation Z_{ijk} will be used to represent this new set of multivariate time series and the notation x_s, y_s and z_s will be used to represent the time series of the x, y and z coordinates of sensor s , where it will be made clear from the context if referring to a general ijk th trial or a specific trial.

Comparison of Z_{ijk} with the PCs of Y_{ijk}

Having transformed the data from Y_{ijk} into an easy to interpret form, Z_{ijk} , as described in section 7.3.3, it is now of interest to discover if this new form supports the basic interpretations of the PCs of Y_{ijk} suggested in section 7.3.2.

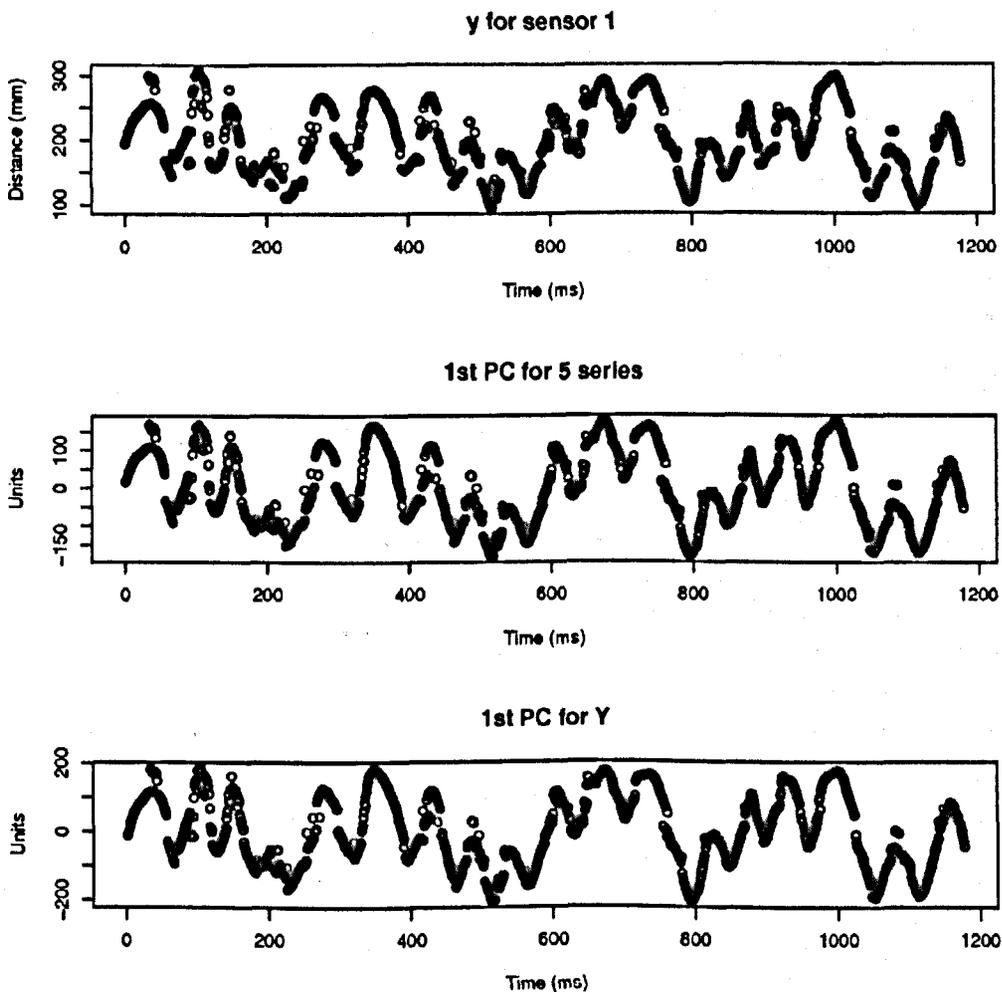


Fig. 7.10: A comparison of y_1 , with the 1st PC of the subset of $Y(t)$ and the 1st PC of $Y(t)$.

Let $Y(t)$, $Y_2(t)$ and $Y_3(t)$ be the three specific sets of data studied in section 7.3.2 and let $Z(t)$, $Z_2(t)$ and $Z_3(t)$ be the corresponding transformed sets of data. It was suggested that the 1st PC of each was related to the 'up and down' movement of sensor 1. In the new Z -notation, this is represented by the y series of sensor 1. Figure 7.10 shows plots of the series y_1 from $Z(t)$, the 1st PC of $Y(t)$ and the

previous approximation of this PC given in equation 7.14. Visually it is difficult to distinguish between the three series, suggesting the 1st PC of $Y(t)$ is indeed attempting to describe the motion of y_1 . The correlation between y_1 and the 1st PC of $Y(t)$ is 0.97. Similar results are found for $Y_2(t)$ and $Y_3(t)$, and the correlations between the 1st PCs of these data sets and the corresponding y_1 series are 0.98 and 0.95 respectively, again supporting the claim that the 1st PC attempts to describe y_1 .

For the 2nd PC of $Y(t)$, the approximation given in equation 7.15 suggested a relationship with the movement of sensor 3. The series y_3 has a correlation of 0.80 with this 2nd PC and a correlation of 0.98 with the approximation given in equation 7.15. For $Y_2(t)$, the 2nd and 3rd PCs both seemed related to the movement of sensor 2, whilst the 4th PC once again seemed related to the movement of sensor 1. Comparison with $Z_2(t)$ shows correlations of 0.77 between x_2 and the 2nd PC and 0.73 between y_2 . The 4th PC had some small correlation with both x_1 and z_1 of 0.32 and 0.27, though generally did not strong correlation with any of the series of $Z_2(t)$. Recall, however, that only the method of accounting for 90% of the total variance suggested retaining this 4th PC whilst the other methods suggested a value of $r \leq 3$.

For the 2nd PC of $Y_3(t)$, the approximation given in equation 7.16 did not lend itself to any simple interpretation. The correlation of this PC with the series y_3 of $Z_3(t)$, however, is 0.95, suggesting this PC is related to the movement of the individuals wrist in the y direction. Since the approximation given in equation 7.16 showed this PC was similar to the weighted sum of $0.4\{1,6\} + 0.6\{2,3\} + 0.7\{3,6\}$, this suggests that the series $\{1,6\}$ may increase as the arm raises, that is that the deodorant stick and the arm may raise at similar times during the application.

PC	x_1	y_1	z_1	x_2	y_2	z_2	x_3	y_3	z_3	y_4	z_4	x_5	y_7	NC	R
1	11	181	22	2	7	1	0	0	0	0	0	0	0	0	0
2	1	0	5	48	28	21	1	28	0	11	5	0	15	61	42
3	1	0	1	9	27	7	1	21	1	11	2	0	2	141	20
4	5	0	6	6	0	4	3	20	1	4	2	2	5	166	0

Tab. 7.5: The number of trials for which the PCs of Y_{ijk} are most correlated with each series of the corresponding Z_{ijk} . NC stands for the number of trials for which the PCs had no correlation greater than 0.7. R stands for the number of these non-correlated PCs which would be retained, taking r as the closest integer to the mean the 6 r s from the methods reported in table 7.4 for each trial.

A comparison of the first 4 PCs of each Y_{ijk} with the corresponding Z_{ijk} is given in table 7.5, which shows the number of trials for which the PCs show correlation of greater than 0.7 with a series from the transformed data. For the 1st PC of each trial, 181 of the 224 trials show the strongest correlation with the series y_1 , with 214 trials related to sensor 1. This suggests once more that the majority of the 1st

PCs are related to the movement of sensor 1, most often in the y direction. The 10 trials for which the 1st PCs were not most correlated with a series of sensor 1 were all most correlated with a series from sensor 2, the other sensor attached to the deodorant stick. Hence the 1st PC for all the trials seems related to the movement of the deodorant stick.

The 2nd, 3rd and 4th PCs can not be consistently interpreted between trials, and in many cases showed no correlation stronger than 0.7 with any of the series of Z_{ijk} . It can be seen, however, that as with the 4th PC of $Y_2(t)$, many of these PCs which show little correlation to any Z_{ijk} would not be retained if the choice of r for each trial was taken as the average r from the 6 methods reported in table 7.4.

The 2nd and 3rd PCs seem most often related to the movement of sensor 2, usually in the x direction, though they are also often related to the movement of sensor 3. The 4th PC, when retained, can be seen to be most often associated with the movement of sensor 3.

7.4 Modelling and Simulation

7.4.1 Grouping

The following procedure for grouping time series is suggested by Alonso and Maharaj [2006], based on the work of Politis and Romano [1994]. Let X and Y be two time series of length n , such that X is generated from the probabilistic model P_X and Y is generated from the probabilistic model P_Y . It is of interest to test if the generating processes are the same for both series, that is, to test the hypothesis

$$H_0 : P_X = P_Y \text{ vs. } H_1 : P_X \neq P_Y \quad (7.20)$$

The method of grouping time series presented here will be useful for several stages of the modelling and simulation of the deodorant stick data. Firstly, if no significant evidence can be found to reject the null hypothesis that two series are generated from the same model, then a single model can be fitted using both of the series. This allows a greater number of observations to be used for each model and fewer models to be required. Since the results of section 7.2 show the effect of individual seems strongly significant, this suggests, for example, that use of one model per individual may suffice. Once the trials have been grouped this claim can be explored.

After fitting models to the data, the grouping method presented here will also be of use to determine if data simulated from the models are significantly different from the original series. In this case, P_X would represent the unknown true generating process and P_Y the fitted model. Hence, this method of grouping can be used to check the fit of the models.

Finally, grouping the data is of interest in its own right. It explores the possibility that several individuals may apply the deodorant stick in a similar manner, adding further insight into the process. If several individuals are found to apply the

deodorant stick in a similar fashion, this discovery would improve the justification of using simulated data in future studies by Unilever, showing that a wider use of these techniques may be expected.

The statistic suggested by Alonso and Maharaj [2006] to carry out the test presented in equation 7.20 is given by

$$T_{n,m} = n \sum_{k=1}^m \left(\hat{R}_X(k) - \hat{R}_Y(k) \right)^2$$

where $\hat{R}_X(k)$ and $\hat{R}_Y(k)$ are the sample ACF of X and Y respectively. An estimate of the distribution of this statistic is found by taking consecutive subsamples of X and Y , $(x_i, x_{i+1}, \dots, x_{i+l})$ and $(y_j, y_{j+1}, \dots, y_{j+l})$, each of length $l > k$, and generating the subsample statistics

$$T_{l,m}^{(i,j)} = l \sum_{k=1}^m \left(\hat{R}_{X_i}(k) - \hat{R}_{Y_j}(k) \right)^2 \quad (7.21)$$

for $i, j \in (1, \dots, n-l+1)$. Using these subsample statistics, the null hypothesis is rejected at the $\alpha\%$ significance level if and only if $T_{n,m} > g_{n,l}(1-\alpha)$, where

$$g_{n,l}(1-\alpha) = \inf \left\{ x : \tilde{G}_{n,l}(x) \geq 1-\alpha \right\}$$

and

$$\tilde{G}_{n,l}(x) = \frac{1}{(n-l+1)^2} \sum_{i=1}^{n-l+1} \sum_{j=1}^{n-l+1} I \left(T_{l,m}^{(i,j)} \leq x \right) \quad (7.22)$$

with $I(E)$ being the indicator function of the event E taking values zero and one.

For the deodorant stick data, the statistic $T_{n,m}$ was found for each series of $Z_{ijk}(t)$ between each pair of trials. Since the trials are not of constant length, for each pair of trials with series lengths n_1 and n_2 , say, with $n_1 > n_2$, then the series of length n_1 was reduced to the length n_2 by omitting an equal number of consecutive observations from the start and finish of the series. The choice of l was then taken as $\lfloor n_2/2 \rfloor$, the integer part of $n_2/2$. The decision to make l dependent on n_2 allows larger l when more observations are available. The choice of m was taken as $m = 10$. This is the same as that taken by Alonso and Maharaj [2006], and brief simulation studies here found varying m to produce little effect. For each pair of trials, the distribution of the $T_{n,m}$ statistic was estimated as in equations 7.21 - 7.22.

Figure 7.11 shows the results of the grouping tests applied to the y_1 series of each pairwise trial at the 5% significance level. Note that within the trials of each individual the majority of tests show no significant difference in the generating processes.

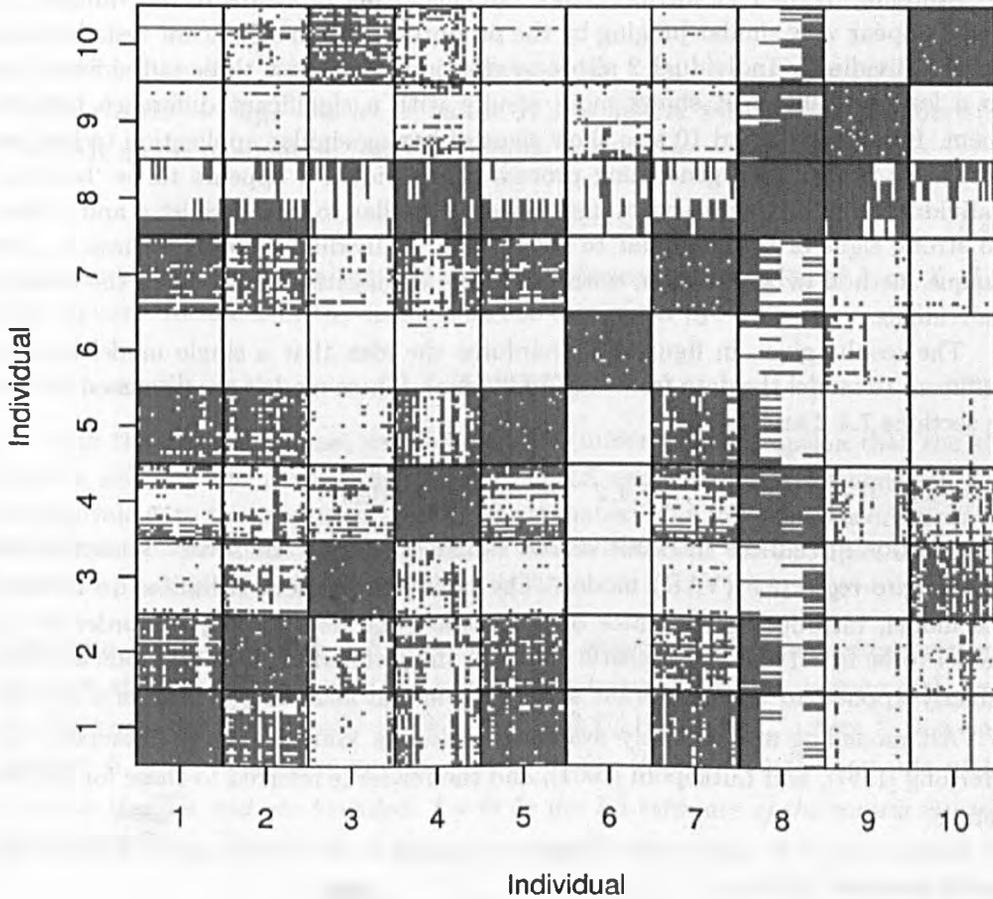


Fig. 7.11: Results of the grouping tests applied to the y_1 series of each pairwise trial. The grey indicates no significant evidence to reject the null hypothesis of the same generating process at the 5% level, whilst the white shows a significant difference between the series was found.

The same is seen for the results of the other series, the figures for which are similar in appearance and thus omitted. This suggests that each individual generates data in the same manner for each trial, meaning that their method of deodorant stick application seems constant between trials. Grouping by product within each individual shows no similar structure, and as with the results of section 7.2, no significant evidence can be found to suggest differences in products from the data.

Studying figure 7.11 further shows the generating processes of individuals 1, 5 and 7 appear very similar judging by the proportion of non-significant tests between these individuals. Individual 2 also appears similar to these three individuals, but to a lesser extent as it shows more results with a significant difference between them. Individuals 3 and 10 also show signs of having similar application techniques with each other. The generating process of individual 8 appears to be 'between' individuals 4 and 9, as it shows signs of being similar to both, whilst 4 and 9 show no strong signs of being similar to each other. Individual 6 seems to have a more unique method of application, with no strong similarities with any of the other 9 individuals.

The results given in figure 7.11 reinforce the idea that a single model may be sufficient to model the data from each individual. These models are discussed further in sections 7.4.2 and 7.4.3.

7.4.2 Vector AR Models

This section introduces the multivariate equivalent of the AR model, known as the *vector auto-regressive* (VAR) model. The review given here is limited to defining the model, introducing estimates of the coefficients, determining the order of the model to be fitted and a diagnostic check on the residuals. These methods are then directly applied to the deodorant stick data in the form $Z_{ijk}(t)$. Several surveys of VAR modelling are currently available, including Watson [1994], Lutkepohl and Breitung [1997], and Lutkepohl [2001], and the reader is referred to these for further details.

Given a set of K time series of length n written in the $K \times n$ matrix Z , the VAR model assumes the form

$$Z(t) - \mu_Z = \Theta(L)(Z(t) - \mu_Z) + \varepsilon(t), \quad (7.23)$$

where $Z(t)$ is the $K \times 1$ column vector of Z corresponding to the time t , μ_Z is the $K \times 1$ row means of Z , $\varepsilon(t)$ is a $K \times 1$ vector of white noise residuals such that $E(\varepsilon(t)\varepsilon(t+u)')=0$ for $u \neq 0$ and $E(\varepsilon(t)\varepsilon(t)')=\Sigma$, where Σ is a $K \times K$ covariance matrix and $\Theta(L)$ is a $K \times K$ matrix of lag coefficients in which the entry of the i th row and the j th column for $i, j \in (1, \dots, K)$ is a lag polynomial of the form

$$\theta_{ij1}L + \theta_{ij2}L^2 + \dots + \theta_{ijp}L^p$$

where L is the Lag operator such that $Ly_t = y_{t-1}$.

Without loss of generality, it is assumed that $\mu_Z = 0$. Now, let X be the $Kp \times n$ matrix, with column vectors

$$X(t) = \begin{pmatrix} Z(t-1) \\ Z(t-2) \\ \vdots \\ Z(t-p) \end{pmatrix} \quad \text{for } t \in (1, \dots, n),$$

let Θ be the $K \times Kp$ matrix of coefficients $[\Theta_1 : \Theta_2 : \dots : \Theta_p]$ where Θ_l is the $K \times K$ matrix of $\{\theta_{ijl}\}$ and let U be the $K \times n$ matrix with column vectors $\varepsilon(t)$. The VAR model of equation 7.23 can then be written as

$$Z = \Theta X + U. \quad (7.24)$$

It can be seen that equation 7.24 is a matrix regression equation of Z on X . The least squares estimates of the coefficients are thus given by

$$\hat{\Theta} = ZX'(XX')^{-1}. \quad (7.25)$$

As in the univariate case, see section 1.6.3, under the assumption that the $\varepsilon(t)$ follow a multivariate normal distribution the LS estimates given in equation 7.25 are equivalent to the maximum likelihood estimates. The following theorem on the asymptotic distribution of these estimates can be found in, for example, Lutkepohl [1991] with reference to Mann and Wald [1943].

Theorem 7.4.1. *Let Z be a $K \times n$ matrix of observed values following the model given in equation 7.24, where U is a $K \times n$ matrix with column vectors $\varepsilon(t)$, such that $E(\varepsilon t) = 0$, $E(\varepsilon(t)\varepsilon(t+u)') = 0$ for $u \neq 0$ and $E(\varepsilon(t)\varepsilon(t)') = \Sigma$, where Σ is a non-singular $K \times K$ covariance matrix. Assume also that all fourth moments of the elements U exist and are bounded. Let $\hat{\Theta}$ be the LS estimate of the matrix Θ given in equation 7.25. Then, as $n \rightarrow \infty$,*

$$\sqrt{n} \text{vec}(\hat{\Theta} - \Theta) \rightarrow_d N(0, \Sigma_{\Theta}),$$

where vec denotes the column stacking operator that stacks the columns of a matrix in a column vector and

$$\Sigma_{\Theta} = \text{plim}_{n \rightarrow \infty} \left(\frac{ZZ'}{n} \right)^{-1} \otimes \Sigma.$$

Remark 7.4.1. As in the univariate case, see section 1.6.3, assumptions must be made for the values of $Z(t)$ for $t \in (-p+1, \dots, 0)$. Two common approaches involve setting these values to zero or shifting the time stamp, such that the first p observed vectors are retained for this purpose and the series length is replaced by $\tilde{n} = n - p$.

Remark 7.4.2. An estimate of Σ can be found from the residuals, $\hat{\varepsilon}(t) = Z - \hat{\Theta}X$, as

$$\hat{\Sigma} = \frac{1}{n} \sum_{t=1}^n \hat{\varepsilon}(t)\hat{\varepsilon}(t)'$$

The results of section 7.4.1 suggested a common set of coefficients between different trials for each individual may fit the deodorant stick data. The extension of equation 7.24 to fit a common model is straight forward. Let Z_1, X_1 and U_1 and Z_2, X_2 and U_2 be the values of Z, X and U from two different trials with common coefficient matrix Θ . Let $Z_c = [Z_1 : Z_2], X_c = [X_1 : X_2]$ and $U_c = [U_1 : U_2]$, be the combined matrixes for the two trials.

Now, note that

$$Z_c = \Theta X_c + U_c$$

and hence the least squares estimates of the coefficients are thus given by

$$\hat{\Theta} = Z_c X_c' (X_c X_c')^{-1}. \quad (7.26)$$

Under the assumption of equality in the covariance matrixes, $\Sigma_1 = \Sigma_2$, the distribution given in Theorem 7.4.1 still applies. The addition of several further trials follows the same argument.

The estimates of the coefficients depend on the order of the model, p . As in the univariate case, this order is generally unknown and must be estimated from the data. One approach is the use of a suitable criterion. The multivariate versions of AIC and BIC, see section 1.6.4, fit the value of p which minimises

$$AIC(p) = \log \left(\det \left(\hat{\Sigma}_p \right) \right) + 2 \frac{pK^2}{n},$$

and

$$BIC(p) = \log \left(\det \left(\hat{\Sigma}_p \right) \right) + \log(n) \frac{pK^2}{n},$$

respectively.

As in the univariate case, see Theorem 1.6.10, the use of AIC asymptotically overestimates the order with positive probability. The use of BIC, however, under general conditions, estimates the order p consistently. See, for example, Paulsen [1984].

Having fitted a VAR(p) model to the data, it is important to check the fit. Since the model assumes $\varepsilon(t)$ to be multivariate white noise, the residuals of the fitted model should be uncorrelated. An inspection of the ACF of each series of residuals may thus be informative, and should show no obvious signs of autocorrelation for $u \neq 0$.

A formal Portmanteau goodness of fit test can be carried out on the residuals to test the null hypothesis $H_0 : E(\varepsilon(t)\varepsilon(t+u)') = 0$, for $u = 1, \dots, h > p$ against $H_1 : \exists u \in (1, \dots, h) : E(\varepsilon(t)\varepsilon(t+u)') \neq 0$. The test statistic has the form

$$Q_h = n \sum_{u=1}^h \text{tr} \left(\hat{C}'_u \hat{C}_0^{-1} \hat{C}_u \hat{C}_0^{-1} \right)$$

where $\hat{C}'_u = n^{-1} \sum_{t=u+1}^n \hat{\varepsilon}(t)\hat{\varepsilon}(t-u)'$. Under the null hypothesis, Q_h has an approximate χ^2 distribution, with $K^2(h-p)$ degrees of freedom, see Ahn [1988].

The adjusted portmanteau statistic is given by

$$Q_h = n^2 \sum_{u=1}^h \frac{1}{n-u} \text{tr} \left(\hat{C}'_u \hat{C}_0^{-1} \hat{C}_u \hat{C}_0^{-1} \right)$$

and has been noted to give potentially superior small sample properties, see Lutkepohl and Kratzig [2004].

Since the test result may depend on the choice of h , it is often a good idea to apply a range of different values. For small values of h , the χ^2 -approximation to the null distribution may be poor. However, a large value of h may result in a loss of power.

The VAR(p) model was fitted to each of the 224 sets of deodorant stick data in the form $Z_{ijk}(t)$. Since the number of coefficients estimated is equal to $15^2 p$ whilst the number of observations for each set of data is equal $15n$, the order was chosen using both AIC and BIC from the range $p \in (0, \dots, \min(c, 20))$, where $c = [(n-1)/15]$, the integer part of $(n-1)/15$, to ensure $15^2 p < 15n$. The smallest value of c is equal to 6, for the case when $n = 105$. The p - value for the adjusted Portmanteau test with $h = 5, 10$ and 25 was recorded for each fitted model for which $h > p$.

Each of the fitted models were then used to simulate 200 new sets of data from each trial, that is, 100 from the model fitted using AIC and a further 100 from the model fitted by BIC. To avoid assumptions on the distributions of the error vectors, $\varepsilon(t)$, the simulated data used a bootstrap approach, reusing the residual vectors, $\hat{\varepsilon}(t)$, of the fitted models in a random order. The simulated series were compared to the actual data using the grouping method described in section 7.4.1.

For each simulated set, $\tilde{Z}_{ijk}(t)$, the data was transformed into the form $\tilde{Y}_{ijk}(t)$ using the standard formula for finding the distance, d , between two points (x_a, y_a, z_a) and (x_b, y_b, z_b) as

$$d = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}. \quad (7.27)$$

PCA was then carried out on the simulated $\tilde{Y}_{ijk}(t)$ and the results compared to those given in section 7.3.2 for the actual data.

The results for the trial $Z_{111}(t)$, referred to as $Z(t)$, will be presented first. This is the data previously discussed in section 7.3. Since the series length of $Z(t)$ is $n = 1178$, this gives a value of $c = \lfloor 1177/15 \rfloor = 78$ and hence the order for the VAR(p) models were chosen using AIC and BIC from the range $(0, \dots, 20)$.

The use of AIC selected a VAR(4) model for the data. This resulted in $15^2 \times 4 = 900$ parameters to be estimated, and thus the model is not written out in full. The parameters were estimated via the LS method given in equation 7.25. Figure 7.12 f) shows the acf of the residuals for the model. Although a small correlation can be seen at lag 5, the adjusted Portmanteau test gave p-values of 0.0769, 0.2077 and 0.3669 for $h = 5, 10$ and 25 respectively. Hence the null hypothesis of uncorrelated residuals is not rejected at the 5% level, and the fitted model seems to account for the autocorrelation structure.

The fitted VAR(4) model was then used to simulate a new set of $\tilde{Z}(t)$ data by resampling the residual vectors $\hat{\varepsilon}(t)$ in a random order for use as the error terms. A comparison of the actual and simulated series of y_1 can be seen in figure 7.12. Comparison of figures 7.12 b) and 7.12 d) shows the ACFs appear similar, although the simulated series seems to possibly decay quicker.

The use of BIC selected a VAR(2) model for the data. The plots for this model are similar in appearance to those for the AIC model and thus omitted. The adjusted Portmanteau test gave p-values of 0.0077, 0.0862 and 0.1769 for $h = 5, 10$ and 25 respectively. It can be seen that when $h = 5$, the null hypothesis of uncorrelated residuals is rejected even at the 1% significance level, and the VAR(2) model may therefore be insufficient to capture full autocorrelation structure of the data.

In order to further evaluate the fit of the two models to the data, 1000 more sets of simulated data were generated from each, resampling the residuals in a random order each time to create the error terms for the simulations. Each simulated series was then compared with the corresponding actual series using the grouping method presented in section 7.4.1, with l taken as 500 and the choice of m taken as 10. The proportions of these simulations which showed significant evidence to reject the null hypothesis of having the same generating process as the original data are shown in table 7.6.

The results agree with those suggested by the Portmanteau statistics, in that the proportions for the AIC model, other than for x_5 at the 1% level, are all within the proportions which would be expected by chance. Hence, no significant evidence is present to suggest the VAR(4) model fitted using AIC is different from the true generating process of the data. However, the model fitted using BIC shows significant evidence to reject this claim for the series z_2, x_3, z_3 and z_4 , although no significant evidence to suggest the VAR(2) model differs from the true generating process for the remaining 11 series. This suggests an order of $p = 2$ is sufficient for the majority of the series, whilst a higher order is required to capture the properties of the remaining series.

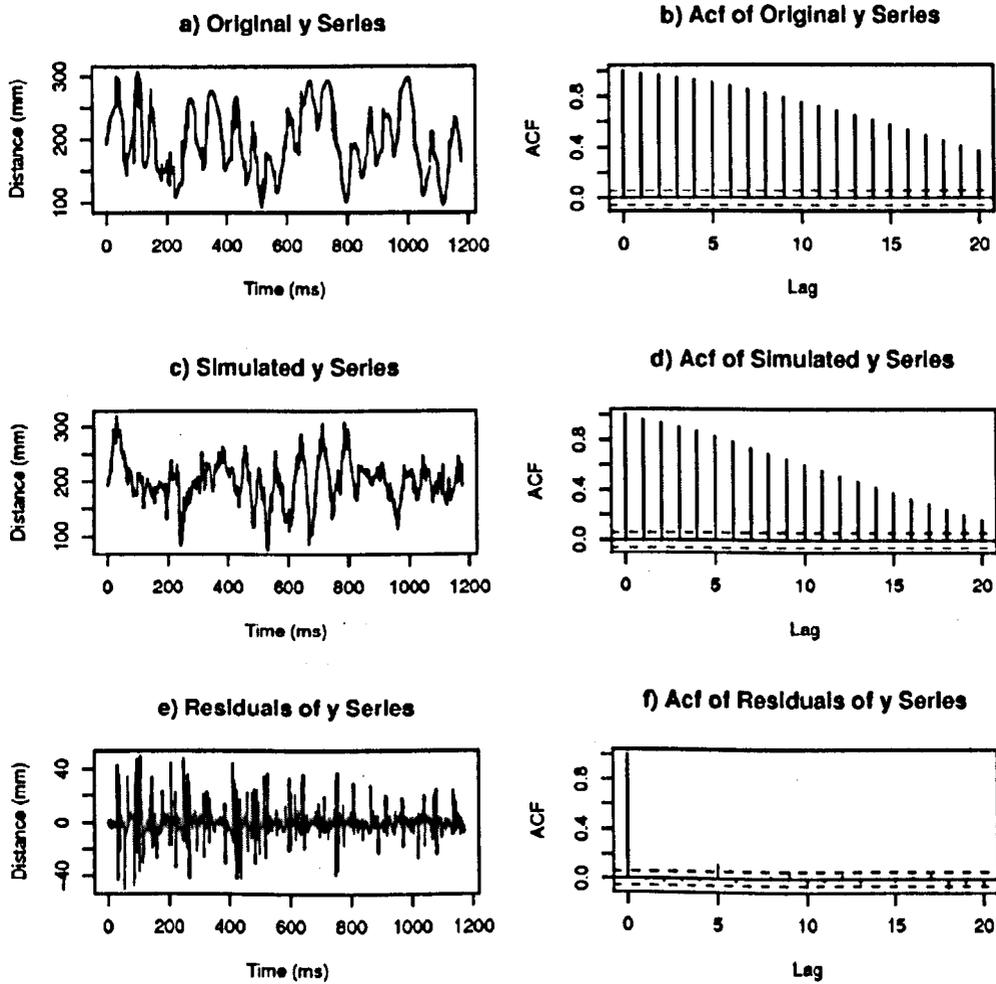


Fig. 7.12: Comparison of the y_1 series of $Z(t)$ with the corresponding series simulated from the VAR(4) model fitted by AIC.

Series	Significance Level					
	VAR(4) using AIC			VAR(2) using BIC		
	10%	5%	1%	10%	5%	1%
x_1	0.00	0.00	0.00	0.01	0.00	0.00
y_1	0.00	0.00	0.00	0.00	0.00	0.00
z_1	0.00	0.00	0.00	0.00	0.00	0.00
x_2	0.01	0.00	0.00	0.01	0.00	0.00
y_2	0.01	0.01	0.00	0.00	0.00	0.00
z_2	0.06	0.01	0.01	0.12	0.05	0.04
x_3	0.01	0.01	0.00	0.14	0.12	0.10
y_3	0.01	0.01	0.00	0.00	0.00	0.00
z_3	0.01	0.00	0.00	0.30	0.13	0.06
x_4	0.00	0.00	0.00	0.00	0.00	0.00
y_4	0.00	0.00	0.00	0.01	0.00	0.00
z_4	0.01	0.01	0.00	0.26	0.13	0.07
x_5	0.09	0.03	0.02	0.00	0.00	0.00
y_5	0.00	0.00	0.00	0.00	0.00	0.00
y_7	0.00	0.00	0.00	0.00	0.00	0.00

Tab. 7.6: Proportion of 100 simulations of $Z(t)$ which reject the null hypothesis given in equation 7.20 at each significance level.

From each of the simulated $\tilde{Z}(t)$, a set of $\tilde{Y}(t)$ was created by transforming the data using equation 7.27. The PCs of the simulated $\tilde{Y}(t)$ were then found. The 1st PCs of the $\tilde{Y}(t)$ generated using the VAR(4) model accounted for between 68% and 81% of the total variance, whilst the 1st PCs of the $\tilde{Y}(t)$ generated using the VAR(2) model generally accounted for a smaller proportion of between 61% and 76%. The 1st PC for the actual data, as discussed in section 7.3.2, accounted for 72% of the total variance, hence the 1st PCs of both models account for a similar proportion as with the original data.

Following the arguments of section 7.3.2, the first PCs of the subset of 5 series involving sensor 1 were considered as an approximation to the 1st PC of the full data sets. The mean equation for the coefficients of these PCs for the 1000 simulations from the VAR(4) model is given by

$$P = -0.38\{1,3\} - 0.45\{1,4\} + 0.43\{1,5\} + 0.49\{1,6\} - 0.47\{1,7\},$$

(0.09) (0.05) (0.07) (0.08) (0.21)

where the figures in brackets show the standard deviations of each coefficient. Comparing this with equation 7.14 shows the 1st PC of this subset of series to be of a similar form to that of the original data. The correlations between the 1st PC

of the subset and the 1st PC of the full data sets were in the range 0.985 and 0.998 with a mean of 0.992. The correlations between the 1st PC of the full data sets and the corresponding simulated y_1 series were in the range 0.952 and 0.975 with a mean of 0.966.

The results for the VAR(2) model are similar, with mean equation for the coefficients of the subset PCs for the 1000 simulations given by

$$P = \begin{matrix} -0.32\{1,3\} & - & 0.46\{1,4\} & + & 0.47\{1,5\} & + & 0.51\{1,6\} & - & 0.43\{1,7\}, \\ (0.11) & & (0.06) & & (0.06) & & (0.07) & & (0.13) \end{matrix}$$

where the figures in brackets show the standard deviations of each coefficient. The correlations between the 1st PC of the subset and the 1st PC of the full data sets were in the range 0.986 and 0.998 with a mean of 0.993. The correlations between the 1st PC of the full data sets and the corresponding simulated y_1 series were in the range 0.951 and 0.978 with a mean of 0.963. The actual data had corresponding correlations of 0.995 and 0.970. The interpretation of the simulated PCs thus appears equivalent to that of the actual data.

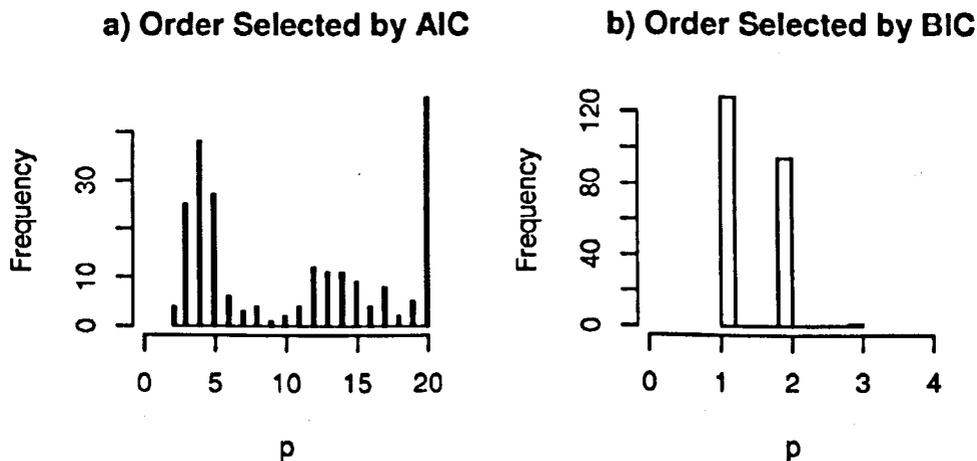


Fig. 7.13: Histograms of the orders selected for the VAR(p) models for the 224 trials using AIC and BIC.

The results are now reported for the remaining trials. Figure 7.13 shows the histograms of the orders of p for the VAR(p) models fitted to the 224 trials using AIC and BIC. It can be seen that the selected order appears to vary greatly between the two criteria. The order selected by BIC has a value of 1 or 2 for almost all the trials, with a maximum fitted order of 3. The use of AIC, on the other hand, has a minimum order fitted of $p = 2$, with a large proportion of the trials fitting a VAR(20) model, the upper limit of the range of fitted values.

For each fitted model, 100 simulated sets of \tilde{Z}_{ijk} were generated by resampling the residuals of each model, and these simulated \tilde{Z}_{ijk} were then transformed into \tilde{Y}_{ijk} using equation 7.27.

AIC	Separately				Cumulatively			
	Min	Max	Mean	SD	Min	Max	Mean	SD
1st PC	0.344	0.999	0.788	0.186	0.344	0.999	0.788	0.186
2nd PC	0.001	0.431	0.117	0.104	0.656	1.000	0.906	0.095
3rd PC	0.000	0.172	0.042	0.040	0.753	1.000	0.947	0.060
4th PC	0.000	0.082	0.018	0.020	0.835	1.000	0.965	0.040
BIC	Min	Max	Mean	SD	Min	Max	Mean	SD
1st PC	0.339	0.958	0.687	0.144	0.339	0.958	0.687	0.144
2nd PC	0.017	0.393	0.170	0.091	0.577	0.975	0.857	0.079
3rd PC	0.011	0.193	0.061	0.038	0.747	0.987	0.918	0.049
4th PC	0.004	0.084	0.027	0.016	0.832	0.991	0.945	0.032

Tab. 7.7: Proportion of variance explained by PCs for the simulated \tilde{Y}_{ijk} .

Table 7.7 shows the proportion of variance explained by PCs for the simulated \tilde{Y}_{ijk} . Comparing the results with those given in table 7.3 for the actual, the 1st PCs of the models fitted using AIC appear to explain more of the variability, whilst those of the BIC models explain less. The standard deviations of the amount of variance explained is larger for the models using AIC than BIC, though both are larger than those reported for the actual data.

The mean equation for the coefficients of the sensor 1 subset PCs for the 1000 simulations are given by

$$\hat{p}_a = 0.52\{1,3\} + 0.49\{1,4\} - 0.40\{1,5\} - 0.41\{1,6\} + 0.27\{1,7\}$$

(0.18) (0.11) (0.10) (0.11) (0.15)

for the models fitted by AIC and

$$\hat{p}_b = 0.44\{1,3\} + 0.52\{1,4\} - 0.45\{1,5\} - 0.48\{1,6\} + 0.21\{1,7\}$$

(0.13) (0.06) (0.07) (0.06) (0.14)

for the models fitted by BIC, where the values in brackets are the standard deviations of the estimated coefficients over the 224 trials. Comparison of this to equation 7.17 shows similarity between the simulated series and the actual data, although the variance in the coefficients has significantly increased. Note, the standard deviations are again larger for the models using AIC.

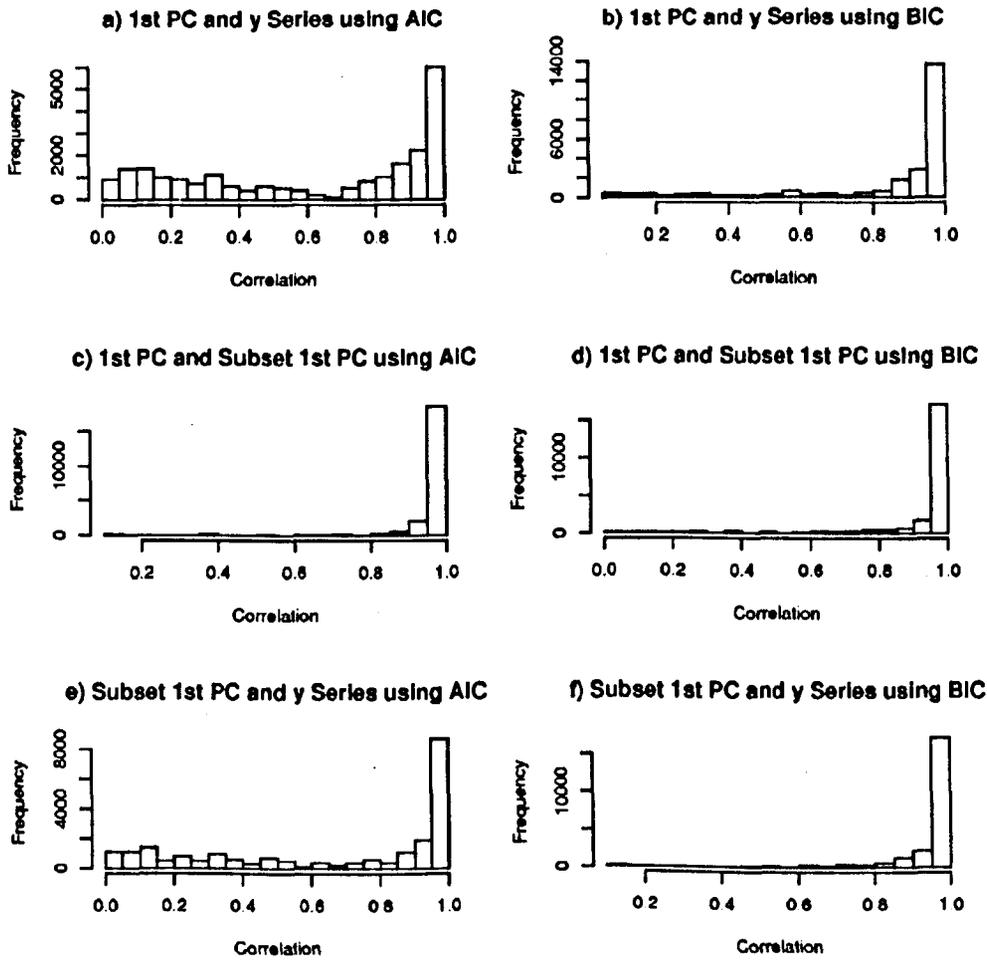


Fig. 7.14: Histograms of the pairwise correlations between the simulated y_1 series, the 1st PC of the simulated $\tilde{Y}_{ijk}(t)$ and the 1st PC of the subset of 5 series from $\tilde{Y}_{ijk}(t)$.

Figure 7.14 shows the histograms of the pairwise correlations between the simulated y_1 series, the 1st PC of the simulated $\tilde{Y}_{ijk}(t)$ and the 1st PC of the subset of 5 series from $\tilde{Y}_{ijk}(t)$. It can be seen that, as with the actual data, the correlation is close to 1 in the vast majority of the simulations. The correlations appear more likely to be smaller for the models fitted using AIC.

As with the results for $Z(t)$, the general results for the 224 trials seem to show that the PCs of the simulated data retain the properties of the actual data, in that the 1st PC accounts for the majority of the variance and can be described by the movement of sensor 1 in the y direction.

Criterion	Individual									
	1	2	3	4	5	6	7	8	9	10
AIC	13	5	14	5	15	4	15	12	14	10
BIC	1	2	1	2	1	2	1	2	2	1

Tab. 7.8: The order selected for the common VAR(p) models fitted to all the trials of each individual.

Now, since the results of section 7.4.1 suggested the trials for each individual to have the same generating process, equation 7.26 is used to estimate the coefficients for a pooled VAR(p) model for each individual. The order selected for each individual using AIC and BIC can be seen in table 7.8. A large difference can once again be seen between the order chosen by AIC and BIC.

For each of the models for each individual, 50 simulated sets of data were generated, taking the error terms in the VAR(p) models from the pooled residuals for each individual. For each of these 50 simulated sets of data, the y_1 series were then compared with each of the y_1 series from the original sets of data for each individual. This resulted in $50 \times \sum_{j=1}^4 K_{ij}$ comparisons being made for the i th individual, with $i \in (1, \dots, 10)$. From table 7.1, it can be seen that 1300 comparisons were made for individuals 1 and 10, 200 comparisons were made for individual 8 and 1200 comparisons were made for the remaining 7 individuals. The proportions of these tests which rejected the null hypothesis of the simulated series having the same generating process as the original series for each individual at the 10%, 5% and 1% significance levels are shown in table 7.9.

It can be seen that only for individuals 2, 4 and 6 does the pooled models fitted using AIC appear adequate for the simulation of the y_1 series at any of the three significance levels. The models fitted using BIC perform somewhat better, although they still show significant evidence to reject the null hypothesis for three individuals at the 10% and 1% significance levels, and 6 individuals at the 5% significance level.

i	AIC			BIC		
	10%	5%	1%	10%	5%	1%
1	0.27	0.23	0.23	0.08	0.08	0.00
2	0.00	0.00	0.00	0.29	0.12	0.00
3	0.46	0.44	0.44	0.04	0.00	0.00
4	0.02	0.00	0.00	0.08	0.08	0.04
5	0.44	0.44	0.44	0.00	0.00	0.00
6	0.02	0.02	0.02	0.13	0.13	0.00
7	0.31	0.31	0.29	0.08	0.04	0.00
8	0.38	0.25	0.12	0.00	0.00	0.00
9	0.23	0.21	0.19	0.25	0.21	0.08
10	0.40	0.40	0.40	0.08	0.08	0.08

Tab. 7.9: The proportion of the y_1 series simulated from the VAR(p) models given in table 7.8 which reject the null hypothesis of having the same generating process as the original trials for the i th individual at the 10%, 5% and 1% significance levels.

7.4.3 B-Splines and Bezier Curves

Due to the generally smooth behaviour of human motion, previous studies of human motion data by Faraway et al [1999], Faraway and Reed [2007] and Faraway et al [2007] suggest the use of fitting smooth functions such as *B-Splines* to the data. The fit of these functions is determined by the set of co-efficients known as the *control points*. These control points are then often linked to covariates of interest, such as the age, height, weight and gender of the individual, by using a regression model. Faraway et al [2007] found these models to be easy to use and interpret, whilst producing realistic simulations and acceptable levels of error for their purpose. Their data consisted of hand trajectories produced whilst the individuals moved an object from a fixed starting point to a given end point in a single motion. They used a special form of B-Splines, known as *Bezier* curves and found four control points were sufficient, with the first and last defined by the end points and only the interior two control points requiring modelling.

This method is not, however, directly applicable to the data studied here for several reasons. First, since no additional information about the individuals or products are available, the model may lack necessary detail. More importantly, however, is that since the deodorant stick application process lacks distinct goals in terms of movement, the control points being modelled may not be describing the same part of the movement between trials. For example, if one trial consists of three up and down motions, whilst another consists of four, extra between trial variability will be introduced into the control points as they will no longer be in alignment. Finally, since the application process is generally more complex than the motions given in these previous studies more control points will be required to give good approximations to the actual data. It would be desirable if simulated control points retained

any autocorrelations found in this series of control points estimated from the actual data.

Taking these differences into account, it is suggested that a time series model such as a VAR process is instead fitted to estimated control points. This section introduces both B-Splines and Bezier functions and describes methods of using these to model the deodorant stick data. Application of these methods is then carried out.

Let $t_i \in [0, 1]$, be a series of $m + 1$ values such that $t_0 \leq t_1 \leq \dots \leq t_m$. These points are referred to as *knots*. The curve, $C(t)$, described by a B-Spline basis can be written as

$$C(t) = \sum_{i=0}^{m-n-1} P_i b_{i,n}(t),$$

where P_i are the control points and $b_{i,n}(t)$ are B-Splines of degree n obtained from the recursive equations

$$b_{i,0}(t) = \begin{cases} 1 & \text{if } t_i \leq t \leq t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$b_{i,n}(t) = \frac{t - t_i}{t_{i+n} - t_i} b_{i,n-1}(t) + \frac{t_{i+n+1} - t}{t_{i+n+1} - t_{i+1}} b_{i+1,n-1}(t),$$

where the t_i , $i \in (0, \dots, m)$ are the knots. Note that $i + n + 1$ can not exceed m , since t_{i+n+1} is not defined for $i + n + 1 > m$. This limits both the number of control points and the degree of the B-Splines, n . Note also that $b_{i,n}(t) = 0$ for $t < t_i$ and $t > t_{i+n+1}$, for all $i \in (0, \dots, m - n - 1)$. This allows improved localised fitting of the curve.

For the case in which the knots are equidistant, the B-Splines are known as uniform B-Splines, and are otherwise referred to as non-uniform. For uniform B-Splines, the basis B-Splines for a given degree n are shifted copies of each other. In this case, a non-recursive definition for the $m - n$ basis B-Splines can be given as

$$b_{i,n}(t) = b_n(t - t_i), \quad i \in (0, \dots, m - n - 1)$$

with

$$b_n(t) = \frac{n+1}{n} \sum_{i=0}^{n+1} (t - t_i)_+^n w_{i,n}$$

and

$$w_{i,n} = \prod_{j=0, j \neq i}^{n+1} \frac{1}{t_j - t_i}$$

where

$$(t - t_i)_+^n = \begin{cases} (t - t_i)^n & \text{if } t - t_i \geq 0 \\ 0 & \text{if } t - t_i < 0 \end{cases}$$

is the truncated power function.

Bezier curves of a degree d are a special case of B-Splines defined by taking $m = 2d + 1$, with the first $d + 1$ knots as $t_0 = t_1 = \dots = t_d = 0$ and the remaining knots as $t_{d+1} = \dots = t_m = 1$. A Bezier curve, $C(t)$, is defined by

$$C(t) = \sum_{i=0}^d P_i B_i^d(t) \quad (7.28)$$

where P_i are once again the control points and $B_i^d(t)$ are the Bezier basis functions of degree d , known as Bernstein polynomials, defined by

$$B_i^d(t) = \binom{d}{i} t^i (1 - t)^{d-i}.$$

Unlike the general B-Splines, all of the Bezier basis functions are supported on the whole interval $[0,1]$. This implies Bezier curves lack the local fitting and numerical stability properties that can be found with other forms of B-spline. Faraway et al [2007] therefore suggest that Bezier curves are not suitable for complex movement such as can be seen in the deodorant stick data here and suggests use of either more general B-Splines or piecewise Bezier functions, such that the original curve is divided into segments and a separate Bezier curve fitted to each.

Applying these curves to the deodorant stick data in the form $Z(t)$, the timestamps, $t \in (m_2, \dots, m_3)$, must first be transformed to $t \in (0, \dots, 1)$. This is done in a linear fashion, such that the new timestamps are still equidistant and the original ordering remains. The B-Spline model then assumes that

$$Z(t) = \sum_{i=0}^{m-n-1} P_i b_{i,n}(t) + \Sigma(t), \quad (7.29)$$

where P_i are the 15×1 control point vectors and $\Sigma(t)$ is the error term. The piecewise Bezier model assumes that

$$Z(t) = \sum_{i=0}^s C_{i,d}(t) + \Sigma(t), \quad (7.30)$$

where the $C_{i,d}(t)$ are Bezier curves defined by equation 7.28 for the segment between t_i and t_{i+1} and zero otherwise and $\Sigma(t)$ is once again the error term. To ensure the segments are connected, the last control point vector of $C_{i,d}(t)$ must equal the first control point vector of $C_{i+1,d}(t)$ for $i \in (0, \dots, s-1)$. Note, however, that this connection need not be smooth.

Fitting these curves requires several choices to be made. The B-Spline model requires a choice for the number of knots, $m+1$, the location of these knots and the degree of the B-Splines, n . Similarly, the piecewise Bezier model requires the number of segments, the location of the breaks and the degree d of each Bezier curve. Once these choices have been made, estimates of P_i can be made using standard linear regression techniques. It should be noted that the location of the knots and break points need not correspond to the locations of observed values.

For simplicity it is assumed that the degrees of the B-Splines and Bezier curves to be fitted are constant for all trials. For B-Splines the degree is related to the smoothness of the curve. If a curve $C(t)$ is produced from a B-Spline basis of degree $n > 1$, only the derivatives up to and including the $(n-2)$ th derivative will be smooth curves. However, since $b_{i,n}(t) > 0$ for $t_i < t < t_{i+n+1}$ for all $i \in (0, \dots, m-n-1)$, larger degrees reduce the localised fit of the curves. Since the deodorant stick data in the form Z_{ijk} is measuring the distances travelled by each sensor in each direction, it is foreseeable that the first and second derivatives may be of future interest as they measure the velocity and acceleration of the sensors. Although these derivatives are not considered here, the choice of $n = 4$ is made to ensure these derivatives would be smooth if required.

For a Bezier curve with degree d , the number of control points is equal to $d+1$. Larger values of d can provide a closer fit to the curve, but do so with an increased number of control points required and a reduction in stability. Although the smoothness of the individual segment Bezier curves are again related to the degree, the use of piecewise Bezier curves implies even the original curve may not be smooth, and thus this is of less concern. Faraway et al [2007] suggest use of $d = 3$ as a suitable compromise between fit and stability for human sensor data. Since brief studies also found little to be gained from higher values of d for the deodorant stick data, the choice of $d = 3$ is also made here.

Given the number of knots and break points, if the locations of the knots and break points were allowed to vary, in addition to modelling the control points, a model would also be required for these locations. Possible interactions of these locations with the control points would cause increased difficulty. The locations of the knots and break points are therefore assumed to be equidistant. Although this may not offer the best possible fit for each trial, it may significantly simplify the models

The number of knots and breaks can be chosen through use of a criterion such as AIC or BIC, see section 1.6.4. For a range of values for the number of each, the respective curves are fitted and the sample standard deviations of the residuals, $\hat{\sigma}^2$,

are recorded. The criteria are then of the form

$$C = \log(\hat{\sigma}^2) + k \frac{p}{T},$$

where T is the number of observations, p is the number of knots/breaks and $k = 2$ for AIC and $k = \log(T)$ for BIC.

Having fitted a B-Spline or piecewise Bezier curve, the control points, P_i , are recorded and a model is fitted to them. It is assumed that the autocorrelation structure of these control points may be significant and thus VAR models are fitted, as described in section 7.4.2.

The models are first fitted to the trial corresponding to $i = 1, j = 1, k = 1$, that is the data $Z(t)$ defined in section 7.3.3. For the B-Splines model, the number of B-Splines fitted ranges from 1 to 60 in steps of 1. For each fit the sample variance of the residuals is recorded and used to discover the best fit in terms of minimising AIC and BIC. In this case the two criteria agree and the number of control points fitted is 49 for both. Figure 7.15e) shows the series of the estimated control points vector corresponding to the y series of sensor 1.

A VAR(p) model is fitted to the series of control point vectors in the manner described in section 7.4.2. The series length for which the model is applied however has been reduced to 49 from the original 1178. The use of AIC and BIC are once again in agreement for the choice of $p = 2$ for the VAR(p) model.

In order to simulate a new set of data, the fitted VAR(2) model is first used to simulate a new set of control points. Since the distribution of the error terms is unknown, the simulated control points are generated by resampling the residuals of the fitted model. The length of the simulated series is kept the same as the original, such that 49 control points are simulated. An example of the series corresponding to the y series of sensor 1 for such a simulation is given in figure 7.15f). A comparison of figures 7.15g) and 7.15h) shows the acf for the simulated control points appears very similar to the originally fitted control points.

Having simulated a new set of control points, substitution of these control points into equation 7.29 provides a new simulation of $Z(t)$. Under the assumption that the error term, $\Sigma(t)$, in equation 7.29 is unnecessary and may be largely due to measurement error, this term is set to zero for the simulations. The original y series of sensor 2 for $Z(t)$ is compared with a simulated series in figures 7.15a)-d). As with the control points, the simulated series and the ACF of the simulated series appear similar to the original.

The grouping test presented in section 7.4.1 was applied to the simulated series to test if the generating process was significantly different from the original. The comparison of y_1 with the simulated series shown in figure 7.15b) with $l = 500$ and $m = 10$ had a p -value of 0.1909 and thus did not reject the null hypothesis that the generating processes were the same at the 5% level. The test was repeated for the

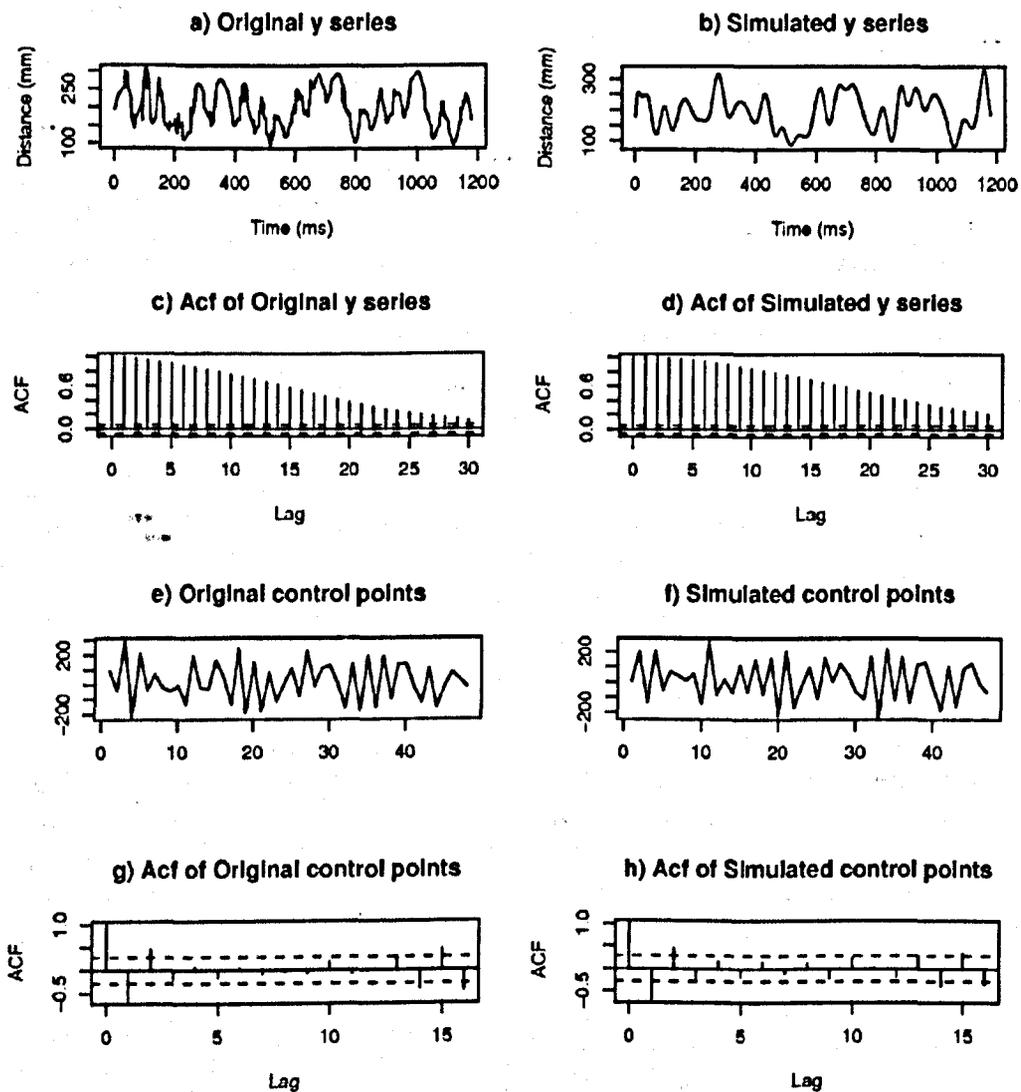


Fig. 7.15: A comparison of the y series of sensor 1 from $Z(t)$ and the corresponding simulated series using the B-Spline model.

remaining 14 series of $Z(t)$ in comparison of their corresponding simulated series and the p -values were all in the range (0.0610, 0.9915). The null hypothesis was thus not rejected for any of the 15 series of $Z(t)$ at the 5% level.

The grouping test was also applied to the original and simulated control points with $l = 25$ and $m = 10$. The p -value for the points corresponding to the y_1 series was 0.7424, and the remaining 14 p -values were in the range (0.0976, 0.8256). The null hypothesis was thus once again not rejected for any of the 15 series of control points at the 5% level.

Series	Significance Level					
	$Z(t)$			P_i		
	10%	5%	1%	10%	5%	1%
x_1	0.37	0.18	0.05	0.03	0.00	0.00
y_1	0.02	0.00	0.00	0.01	0.00	0.00
z_1	0.00	0.00	0.00	0.04	0.00	0.00
x_2	0.04	0.02	0.01	0.01	0.00	0.00
y_2	0.18	0.13	0.10	0.03	0.00	0.00
z_2	0.06	0.01	0.00	0.01	0.00	0.00
x_3	0.00	0.00	0.00	0.00	0.00	0.00
y_3	0.00	0.00	0.00	0.00	0.00	0.00
z_3	0.38	0.33	0.29	0.01	0.00	0.00
x_4	0.00	0.00	0.00	0.06	0.02	0.00
y_4	0.00	0.00	0.00	0.00	0.00	0.00
z_4	0.38	0.30	0.21	0.09	0.03	0.01
x_5	0.00	0.00	0.00	0.02	0.00	0.00
y_5	0.00	0.00	0.00	0.01	0.00	0.00
y_7	0.00	0.00	0.00	0.02	0.00	0.00

Tab. 7.10: Proportion of 1000 simulations of $Z(t)$ and the B-Spline control points which reject the null hypothesis given in equation 7.20 at each significance level.

A further 1000 sets of simulated data were then generated in the same manner from the fitted model and the grouping test was applied to each in comparison with the original data. For the simulated control points l was taken as 25, whilst for the full simulated series l was taken as 500. The choice of m was taken as 10 for both. The proportions of these simulations which showed significant evidence to reject the null hypothesis of being generated from the same process as the original data are shown in table 7.10.

It can be seen that, for the simulated control points, the proportions of simulated series which show significant difference to the original data are all within the proportions which may be expected by chance under the null hypothesis. Hence, no significant evidence is found to reject this model for the control points.

Although this is also true for the majority of the simulated data generated from

the simulated control points, it can be seen that a large proportion of the simulated series of x_1 , y_2 , z_3 and z_4 do show significant evidence to reject the null hypothesis that they are generated by the same process as the original data. This suggests a possible lack of fit for these four series. It should be noted, however, that the results of section 7.3.3 showed the majority of the variability of $Z(t)$ is accounted for by the series y_1 and y_3 , and the model appears to produce a good fit for both of these series.

For the Bezier model, the number of segments fitted ranges from 1 to 30 in steps of 1. As with the B-Splines model, the sample variance of the residuals was recorded for each and used to find the number of segments which minimised AIC and BIC. The use of AIC fitted a model with 25 segments, whilst the use of BIC fitted a model with 21. Since the last control point of the i th segment is set to equal the first of the $(i + 1)$ th segment, the total number of control points for the model fitted by AIC is 76, whilst the number of control points for the model fitted by BIC is 63.

A VAR(p) model is fitted to each of the series of control point vectors in the manner described in section 7.4.2. For the control points of the AIC Bezier model, the use of AIC fits a VAR(4) model. For the control points of the BIC Bezier model, further use of BIC fits a VAR(3) model.

As with the B-Spline model, new sets of data are simulated by first using the VAR(p) models to simulate a new set of control points. Then, having simulated a new set of control points, substitution of these control points into equation 7.30 provides a new simulation of $Z(t)$. The error terms used for simulating new control points from the VAR(p) are once again generated by resampling the residuals of the fitted model and the length of the simulated series is kept the same as the original. The error term, $\Sigma(t)$, in equation 7.30 is again assumed negligible and set to zero for the simulations.

A comparison of the y series of sensor 1 from $Z(t)$ and the corresponding simulated series generated from the Bezier model fitted using BIC can be seen in figure 7.16. The grouping test presented in section 7.4.1 was again applied to the simulated series to test if the generating process was significantly different from the original. The comparison of y_1 with the simulated series shown in figure 7.16b) with $l = 500$ and $m = 10$ had a p -value of 0.2159. The grouping test applied to the original and simulated control points corresponding to the y_1 series with $l = 25$ and $m = 10$ had a p -value of 0.5987. Thus no significant evidence was found to reject the null hypothesis that the generating processes were the same at the 5% level.

The test was repeated for the remaining 14 series of $Z(t)$ in comparison of their corresponding simulated series and the p -values were all in the range (0.0067, 0.9729). The null hypothesis was rejected at the 5% level for the series x_1 , z_3 and z_4 . Note, as seen from table 7.10, these are the same series which the B-Spline model often also failed to adequately simulate. The remaining 14 p -values in the comparison of the original and simulated control points were in the range (0.0686, 0.8861). The null

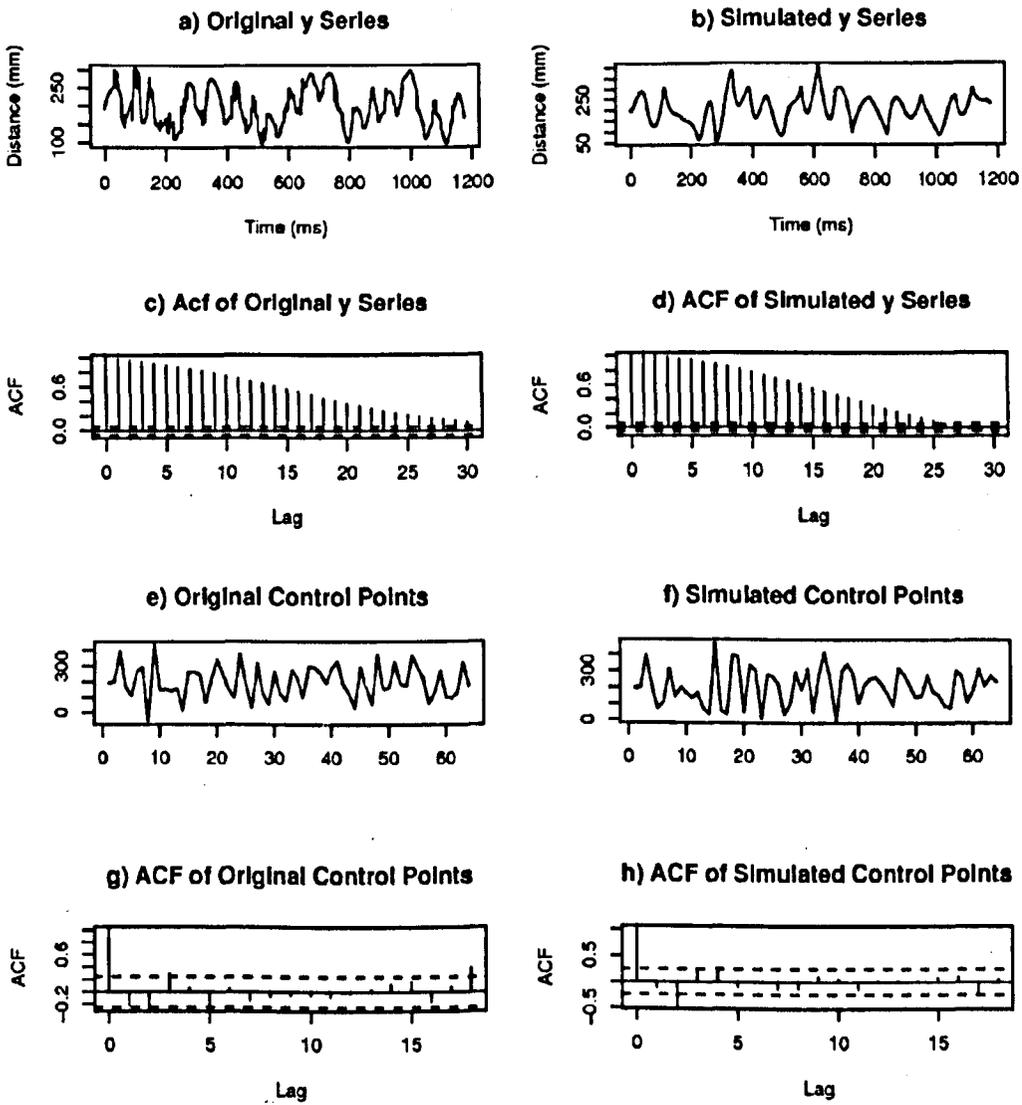


Fig. 7.16: A comparison of the y series of sensor 1 from $Z(t)$ and the corresponding simulated series generated from the Bezier model fitted using BIC.

hypothesis was thus not rejected for any of the 15 series of control points at the 5% level.

A figure containing a comparison of the y_1 and the corresponding simulated series generated from the Bezier model fitted using AIC is omitted since it is similar in appearance to figure 7.16. The comparison test for the simulated y_1 had a p -value of 0.4188, whilst the p -value for the corresponding simulated control points was 0.2755. Thus no significant evidence was found to reject the null hypothesis that the generating processes were the same at the 5% level for the y_1 series of this simulated data.

For the remaining 14 series, however, the p -values for the simulated series were in the range (0, 0.7741) and the p -values for the simulated control points were in the range (0, 0.4392), with significant evidence at the 5% level to reject the null hypothesis for series x_2, z_2, x_3, x_4 and z_4 and 5 of the 15 series of control points. The VAR(4) model fitted to the control points using AIC does not seem to fit the data well. The large number of estimated parameters may have made the model less stable.

Series	Significance Level											
	Fitted by AIC						Fitted by BIC					
	$Z(t)$			P_i			$Z(t)$			P_i		
	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
x_1	0.01	0.01	0.00	0.15	0.00	0.00	0.20	0.03	0.00	0.05	0.04	0.03
y_1	0.26	0.18	0.13	0.13	0.07	0.04	0.01	0.00	0.00	0.02	0.02	0.01
z_1	0.02	0.01	0.01	0.17	0.00	0.00	0.00	0.00	0.00	0.05	0.04	0.03
x_2	0.49	0.18	0.10	0.17	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.01
y_2	0.38	0.22	0.13	0.11	0.03	0.01	0.01	0.01	0.00	0.00	0.00	0.00
z_2	0.53	0.20	0.07	0.17	0.00	0.00	0.08	0.05	0.01	0.01	0.00	0.00
x_3	0.48	0.17	0.06	0.17	0.00	0.00	0.00	0.00	0.00	0.06	0.04	0.04
y_3	0.19	0.15	0.12	0.10	0.07	0.05	0.00	0.00	0.00	0.02	0.01	0.01
z_3	0.00	0.00	0.00	0.17	0.00	0.00	0.51	0.33	0.12	0.03	0.02	0.02
x_4	0.51	0.33	0.09	0.17	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.03
y_4	0.39	0.07	0.01	0.13	0.01	0.00	0.08	0.04	0.01	0.10	0.08	0.07
z_4	0.50	0.24	0.03	0.17	0.00	0.00	0.11	0.07	0.04	0.08	0.05	0.04
x_5	0.10	0.07	0.05	0.14	0.02	0.00	0.00	0.00	0.00	0.07	0.06	0.04
y_5	0.23	0.20	0.17	0.13	0.09	0.06	0.00	0.00	0.00	0.05	0.04	0.03
y_7	0.30	0.16	0.09	0.13	0.02	0.00	0.06	0.04	0.01	0.11	0.08	0.06

Tab. 7.11: Proportion of 1000 simulations of $Z(t)$ and the Bezier control points which reject the null hypothesis given in equation 7.20 at each significance level.

A further 1000 sets of simulated data were then generated from both the AIC and BIC Bezier models. The grouping test was applied to each simulation in comparison with the original data, with the same choices of l and m as previously. The

proportions of the simulations from the AIC and BIC Bezier models which showed significant evidence to reject the null hypothesis of being generated from the same process as the original data are shown in table 7.11.

The proportion of simulated series from the Bezier model fitted using AIC which show significant evidence to reject the null hypothesis are higher than expected by chance for every series except x_1 , z_1 and z_3 . Indeed, for several of the series more than 10% of the simulations rejected the null hypothesis at the 1% significance level, including the simulations of y_1 . The results for the simulated control points from this model show every series having greater than or equal to 10% of the simulations being significantly different from the original estimated control points at the 10% significance level. At the 5% and 1% significance levels, the results for the control points generally improve, however the simulations for the series y_1 , y_3 and y_5 still appear significantly different. This is concerning, since y_1 and y_3 have previously been found in section 7.3.3 to represent the first two PCs of $Y(t)$, accounting for the majority of the variance within the data. These results suggest the AIC Bezier model may not fit the data well.

The proportions of simulations which reject the null hypothesis for the BIC Bezier model are generally smaller than the corresponding values for the AIC Bezier model, with only the simulations for x_1 , z_3 and z_4 having proportions larger than expected. These are three of the four series which also showed a lack of fit for the B-Spline model and, in comparison, the proportions for these series are smaller for the BIC Bezier model. Despite this, the simulated control points for the BIC Bezier model show much larger than expected proportions of the simulations being significant at the 1% level for the majority of the series.

The performances of the BIC Bezier model and the B-Spline model in simulating new sets of $Z(t)$ are comparable. Although the simulations from the BIC Bezier model seem to better capture the original ACF of the series x_1 , y_2 , z_3 and z_4 the series are generally not smooth. The simulations from the B-Spline model, on the other hand, are smooth and can be used to find the first two derivatives if required. The choice of which of these models to use may therefore depend on the practical usage of the simulations.

Having fitted the B-Spline and Bezier models to $Z_{111}(t)$, the models were then fitted to the remaining trials. For each of the 224 trials, two B-Spline models were fitted to Z_{ijk} using both AIC and BIC to choose the number of control points, $\hat{m} - 4$, from $m \in (5, \dots, 64)$. Two Bezier models were also fitted to Z_{ijk} using these criteria to choose the number of segments from $s \in (1, \dots, 30)$. The resulting number of control points and segments fitted can be seen as histograms for the 224 trials in figure 7.17.

For the B-Spline model, the use of AIC finds the modal number of control points required for the trials to be 50, whilst for the Bezier model, the modal number of segments is 24. The distributions of these numbers using BIC is more bimodal, with the higher modes in agreement with the results using AIC. Recall that the B-Spline

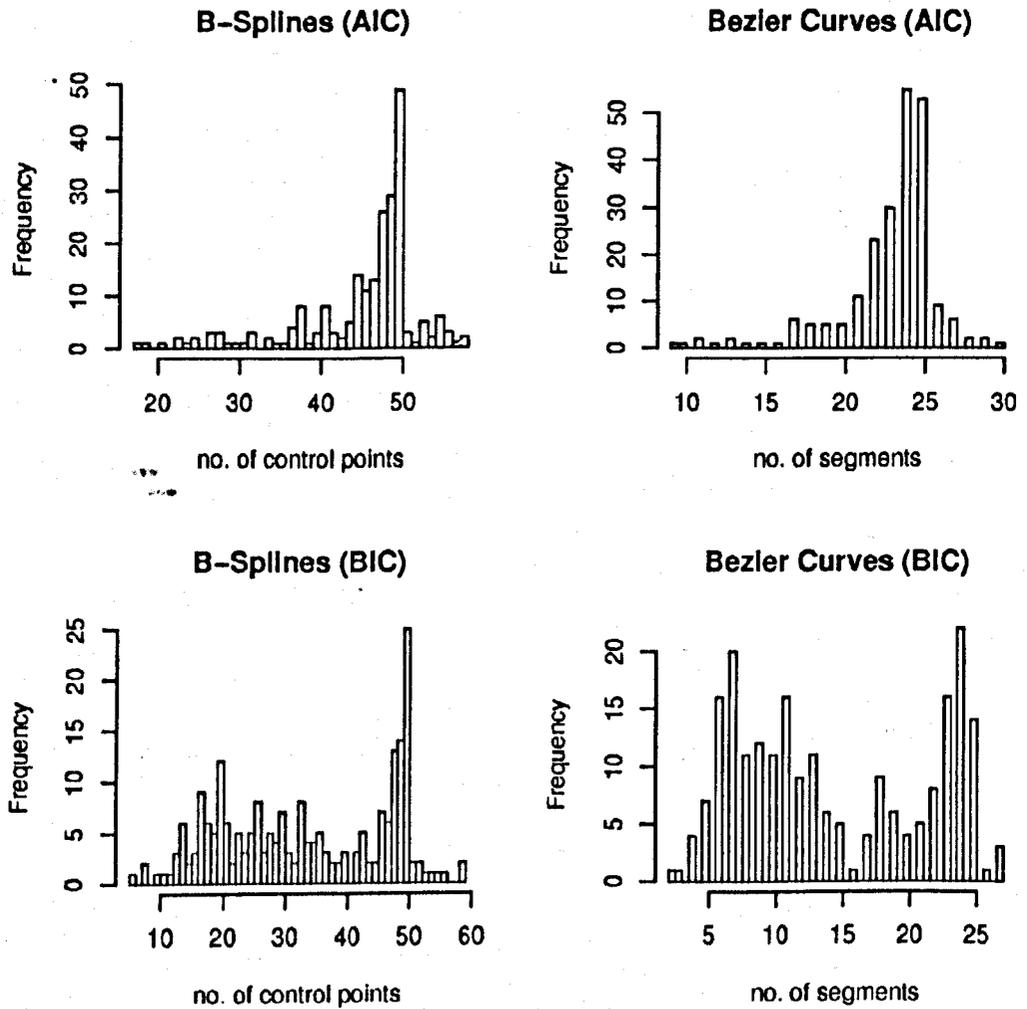


Fig. 7.17: The number of control points and segments fitted to the 224 trials using AIC and BIC.

models fitted for $Z(t)$ had 49 control points using both AIC and BIC and the Bezier models fitted had 21 segments using BIC and 25 segments using AIC. Figure 7.17 shows these values to be fairly typical amongst the trials.

Having found the control points for each model, a VAR(p) model was then fitted to them, with the choice of $p \in (0, \dots, 10)$. For models in which the number of control points had been decided by use of AIC, the order p of the VAR model was also found using AIC. Similarly, where the number of control points had been decided by use of BIC, the order p of the VAR model was also found using BIC. Histograms of the orders chosen can be found in figure 7.18.

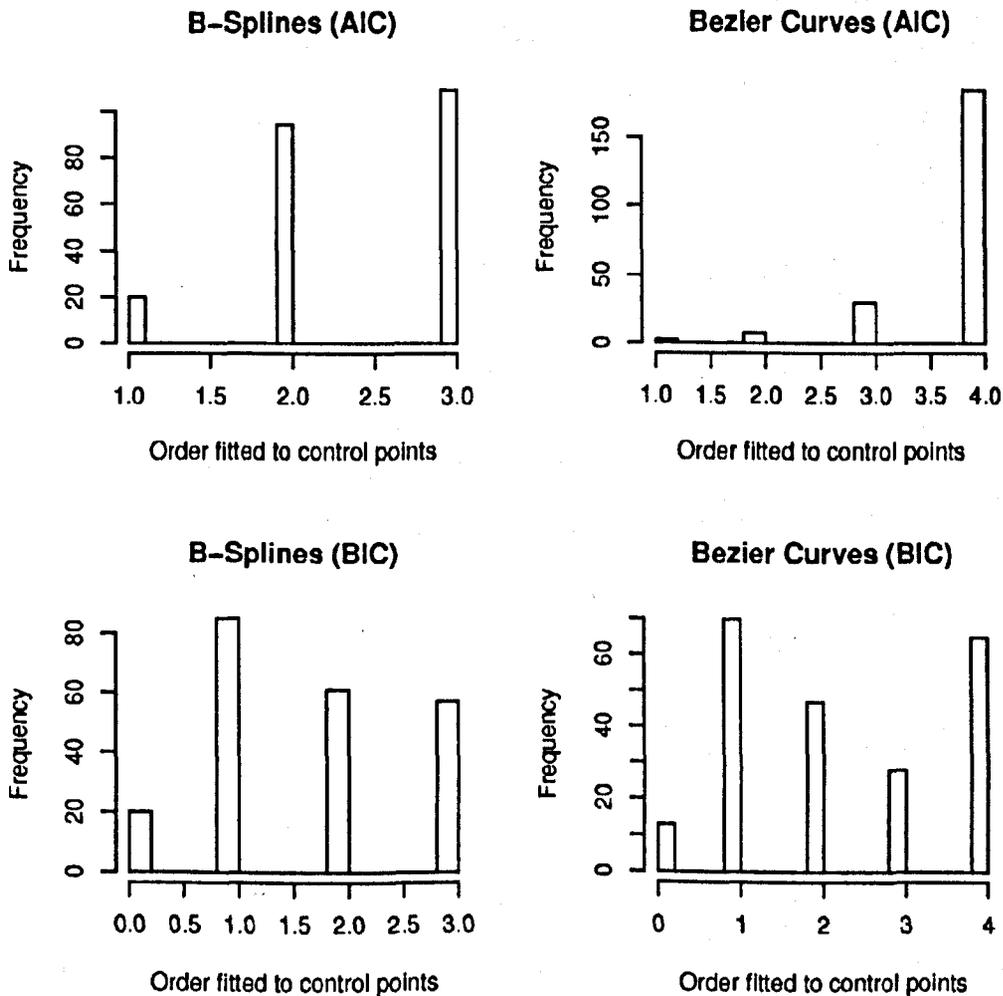


Fig. 7.18: The order of the VAR(p) models fitted to the control points for the 224 trials using AIC and BIC.

It can be seen that the order fitted, p , was less than or equal to 3 for the B-

Spline models and less than or equal 4 for the Bezier models using both AIC and BIC. The use of AIC tended to fit these maximum orders, particularly for the Bezier model, and fitted a minimum order of $p = 1$. The use of BIC, in comparison, fitted a VAR(0) model, that is, multivariate white noise, to the control points of the B-Spline model in 20 trials and the control points of the Bezier model in 12, though generally favoured the fit of a VAR(1) model for both.

As mentioned in section 7.4.2, the results of section 7.4.1 suggest a single model for each individual to be sufficient. Hence, the number of control points for each individual are assumed constant and a single VAR(p) model is fitted to each of the individual's trials. The AIC and BIC are found using the pooled residuals and the pooled VAR(p) models introduced in section 7.4.2 are once again used to fit the number of control points and the order p of the VAR model. The fitted number of control points and the order of the fitted VAR(p) model for each individual can be seen in table 7.12.

	B-Spline Model				Bezier Model			
	AIC		BIC		AIC		BIC	
i	c	p	c	p	c	p	c	p
1	44	2	28	1	67	4	34	2
2	49	3	43	2	73	4	58	3
3	38	2	20	1	67	4	25	1
4	50	3	49	3	76	4	73	4
5	40	2	18	1	61	3	22	1
6	50	3	49	3	76	4	73	4
7	44	2	33	2	70	4	40	2
8	48	2	46	2	73	4	55	3
9	49	3	36	2	70	4	46	3
10	44	2	28	1	70	4	34	2

Tab. 7.12: The number of control points, c , and the order of the VAR(p) model fitted to the control points when fitting a single model to the i th individual.

Note from table 7.12 that both the number of control points and the order of the VAR(p) model fitted to these control points is smaller for the B-Spline model for every individual using both AIC and BIC to fit the model.

For each of the four models for each individual, 50 simulated sets of data were generated, taking the error terms in the VAR models from the pooled residuals for each individual. For each of these 50 simulated sets of data, the y_1 series were then compared with each of the y_1 series from the original sets of data for each individual. This resulted in $50 \times \sum_{j=1}^4 K_{ij}$ comparisons being made for the i th individual, with $i \in (1, \dots, 10)$. From table 7.1, it can be seen that 1300 comparisons were made for individuals 1 and 10, 200 comparisons were made for individual 8 and

1200 comparisons were made for the remaining 7 individuals. The proportions of these tests which rejected the null hypothesis of the simulated series having the same generating process as the original series for each individual at the 10%, 5% and 1% significance levels are shown in table 7.13.

<i>i</i>	B-Spline Model						Bezier Model					
	AIC			BIC			AIC			BIC		
	10%	5%	1%	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	0.08	0.04	0.00	0.11	0.04	0.00	0.10	0.09	0.04	0.03	0.00	0.00
2	0.07	0.07	0.05	0.10	0.04	0.00	0.07	0.05	0.00	0.08	0.04	0.00
3	0.04	0.03	0.01	0.08	0.04	0.00	0.04	0.00	0.00	0.04	0.00	0.00
4	0.03	0.00	0.00	0.02	0.00	0.00	0.03	0.01	0.00	0.08	0.00	0.00
5	0.02	0.00	0.00	0.03	0.01	0.00	0.04	0.00	0.00	0.01	0.00	0.00
6	0.21	0.16	0.12	0.16	0.14	0.09	0.35	0.33	0.31	0.31	0.24	0.14
7	0.10	0.10	0.04	0.12	0.10	0.04	0.10	0.07	0.05	0.08	0.07	0.05
8	0.10	0.00	0.00	0.05	0.00	0.00	0.02	0.00	0.00	0.10	0.05	0.00
9	0.11	0.04	0.00	0.12	0.12	0.04	0.40	0.40	0.37	0.30	0.22	0.15
10	0.08	0.02	0.00	0.11	0.04	0.04	0.28	0.22	0.12	0.03	0.00	0.00

Tab. 7.13: The proportion of the y_1 series simulated from the models given in table 7.12 which reject the null hypothesis of having the same generating process as the original trials for the i th individual at the 10%, 5% and 1% significance levels.

It can be seen that the fit of all four models for individual 6, and to a lesser extent individual 7, seem inadequate, with large proportions of the simulated series showing significant difference to the generating process of the original trials. The simulations from the Bezier models and the B-Spline model fitted using BIC for individual 9 also appear significantly different from the original series. The B-Spline model fitted using AIC appears to generate simulations closer to the original data for individual 9.

The performances of each of the four models appear similar overall. In addition to the difficulties already mentioned, the B-Spline model using AIC appears to lack fit for individual 2, the BIC B-Spline model lacks fit for individual 10 and the AIC Bezier model lacks fit for individuals 1 and 10. All four models appear to provide a good fit for individuals 3,4,5 and 8. Overall, the simulations for at least one of the models have shown no significant difference from the actual data for 8 of the 10 individuals.

Comparison of table 7.13 with tables 7.10 and 7.11 gives an idea of the relative performance of the pooled model with the separated model for individual 1. The proportions of significantly different series have increased for both the B-Spline models and the BIC Bezier model, although the fit of the pooled models are still appear sufficient. The proportions for the AIC Bezier model have actually decreased, although a lack of fit can be seen in both cases.

7.5 Summary

Several key issues have been addressed during the analysis of the deodorant stick data. The data has been transformed into an easy to interpret form with a reduction in data without any loss of information. The key series have then been identified by comparison of this new form to the PCs of the old.

The results of the various tests have shown significant evidence of differences between individuals and suggested common application techniques exist. This thus gives credibility to the use of the deodorant stick data from these 10 individuals to simulate further data for future research at Unilever, as it shows that even between 10 randomly selected individuals similarities between the application techniques can be found, and hence these application techniques may be common in the population as a whole. The study has also found that the current experiment is unsuited to find significant differences between products.

Various models for simulating the data were then fitted. Although the VAR models often accounted for the correlation structure of the original data, they produce simulated series which appear more jagged. In general, the B-Spline and Bezier models have been found to be able to provide simulations similar to the original data in both appearance and structure of ACF, and these methods could be used to produce the simulations desired by Unilever.

BIBLIOGRAPHY

- [1988] S. K. Ahn. Distribution of residual autocovariances in multivariate autoregressive models with structured parameterisation. *Biometrika*. 75, 590-593.
- [1969] H. Akaike. Power Spectrum estimation through autoregressive model fitting. *Ann. Inst. Statist. Math.* 21, 407-419.
- [1974] H. Akaike. A new look at the statistical model identification. *Transactions on Automatic Control* 19 (6): 716-723.
- [1977] H. Akaike. On entropy maximisation principle. *Applications of Statistics*. 27-41.
- [2006] A. M. Alonso and E. A. Maharaj. Comparison of time series using subsampling. *Computational Statistics and Data Analysis*. 50, 10, 2589-2599.
- [2000] J. F. Alves and V. Araujo. Random perturbations of non-uniformly expanding maps. *Arxiv preprint math.DS/0011158*.
- [2004] J. F. Alves, C. Bonatti and M. Viana. SRB measures for partially hyperbolic systems whose central direction is mostly expanding. *The Theory of Chaotic Attractors*. 140, 443.
- [2002] J. F. Alves and M. Viana. Statistical stability for robust classes of maps with non-uniform expansion. *Ergodic Theory and Dynamical Systems*. 22, 1, 1-32.
- [1998] G. S. Ammar. Classical foundations of algorithms for solving positive definite Toeplitz equations. *Calcolo. A Quarterly on Numerical Analysis and Theory of Computation*. 33, 99-113.
- [1942] R. L. Anderson. Distribution of the serial correlation coefficient. *Ann. Math. Stat.* 13,1.
- [1971] T. W. Anderson. *The Statistical Analysis of Time Series*. New York: John Wiley.
- [1967] D. V. Anosov. Geodesic Flows on a Compact Riemann Manifold of Negative Geodesic Curvature. *Proc. Steklov Inst. Math* 90,1.

- [2000] V. Baladi. Positive transfer operators and decay of correlations. *World Scientific*. 16.
- [1997] V. Balakrishnan, G. Nicolis and C. Nicolis. Recurrence time statistics in chaotic dynamics. I. Discrete time maps. *Journal of Statistical Physics*, 86(1/2):191- 212.
- [2001] V. Balakrishnan, G. Nicolis and C. Nicolis. Recurrence time statistics in chaotic dynamics: multiple recurrences in intermittent chaos. *Stochastics and Dynamics*, 1(3):345 - 359.
- [2001] V. Balakrishnan and M. Theunissen. Power-law tails and limit laws for recurrence time distributions. *Stochastics and Dynamics*, 1(3):339-343.
- [2003] J. M. Bardet, G. Lang, G. Oppenheim, A. Phillippe, S. Stoev and M. S. Taquq. Semi-parametric estimation of the long-range dependence parameter: A survey. *Theory and Applications of Long-Range Dependence*, 557-577.
- [1946] M. S. Bartlett. Statistics On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society* 1(8). 27-41.
- [1948] M. S. Bartlett. Statistical estimation of supply and demand relations. *Econometrica* 16. 323-329.
- [1950] M. S. Bartlett. Tests of significance in facto analysis. *Br. J. Psych. Stat.* 3, 77-85.
- [1987] K. I. Beltrao and P. Bloomfield. Determining the bandwidth of a kernel spectrum estimate. *J. Time Ser. Anal.* 8, 21-38.
- [1996] P. M. Bentler and K. H. Yuan. Test of linear trend in eigenvalues of a covariance matrix with application to data analysis. *Brit. J. Math. Statist. Psychol.* 49, 299-312.
- [1998] P. M. Bentler and K. H. Yuan. Test of linear trend in the smallest eigenvalues of the correlation matrix. *Psychometrika.* 63, 131-144.
- [1993] J. Beran. Fitting Long-memory models by generalized linear regression. *Biometrika* 80, 4, 817-822.
- [1994] J. Beran. Statistics for Long Memory Processes. *Chapman and Hall, London*.
- [1998] J. Beran, R. J. Bhansali and D. Ocker. On Unified model Selection for stationary and non-stationary short- and long-memory autoregressive processes. *Biometrika.* 85, 921-934.

-
- [1974] K. N. Berk. Consistent autoregressive spectral estimates. *Ann. Stat.* 2, 489-502.
- [1992a] L. M. Berliner. Statistics, Probability and Chaos. *Statistical Science* 1(7), 69-90.
- [1992b] L. M. Berliner. [Statistics, Probability and Chaos.]: Rejoinder *Statistical Science* 1(7), 114-122.
- [1997] R. J. Bhansali. Robustness of the autoregressive spectral estimate for linear processes with infinite variance. *Journal of Time Series Analysis* 18, 3, 213-229.
- [1977] R. J. Bhansali and D. Y. Downham. Some properties of the order of an autoregressive model selected by an extension of Akaike's FPE criterion. *Biometrika* 64, 547-51.
- [2006] R. J. Bhansali, M. P. Giraitis and P. S. Kokoszka. Estimation of the memory parameter by fitting fractionally differenced autoregressive models. *Journal of Multivariate Analysis*, 97, 2101-2130.
- [2005] R. J. Bhansali, M. P. Holland and P. S. Kokoszka. Chaotic maps with slowly decaying correlations and intermittency. *Fields Institute Communications*, 44, 99-126.
- [2006] R. J. Bhansali, M. P. Holland and P. S. Kokoszka. Intermittency, long-memory and financial returns. *Long Memory in Economics*. G. Teyssiere and A. Kirman (eds), 39-68.
- [2008b] R. J. Bhansali and M. P. Holland. Frequency Analysis of Chaotic Intermittency Maps with Slowly Decaying Correlations.
- [2001a] R. J. Bhansali and P. S. Kokoszka. Estimation of the long-memory parameter: a review of recent developments and an extension. *Proceedings of the Symposium on Inference for Stochastic Processes* I. V. Basawa, C. C. Heyde and R. L. Taylor, eds. 125-150.
- [2001b] R. J. Bhansali and P. S. Kokoszka. Prediction of Long-Memory Time Series: An Overview. *Estadística* 160, 41-96.
- [2002] R. J. Bhansali and P. S. Kokoszka. Computation of the forecast coefficients for multistep prediction of long-range dependent time series. *International Journal of Forecasting* 18, 181-206.
- [1986] P. Billingsley. Probability and Measure. *Wiley-Interscience, New York*.
- [1987] N. H. Bingham, C. M. Goldie, J. L. Teugels. Regular Variation. *Encyclopaedia of Mathematics and its Applications* 27.

- [2002] L. Bisgalia. Model selection for long-memory models. *Quaderni di Statistica*. 4, 33-49.
- [2002] L. Bisgalia and S. Bordignon. Mean square prediction error for long-memory processes. *Statistical Papers*. 43, 161-175.
- [1973] P. Bloomfield. An exponential model for the spectrum of a scalar time series. *Biometrika* 60, 217-226.
- [1976] G. E. P. Box and G. M. Jenkins. Time Series Analysis: forecasting and control. *Holden-Day Series in Time Series Analysis*.
- [1970] G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Jour. Amer. Stat. Assoc.* 64.
- [1970] R. Bowen. Markov Partitions for Axiom A Diffeomorphisms. *American Journal of Mathematics* 92, 725.
- [1975] R. Bowen and D. Ruelle. Ergodic Theory of Axiom A Flows. *Inventiones Mathematicae* 79, 181-202.
- [1972] D. R. Brillinger. The Analysis of Time Series Collected in an Experimental Design. *Multivariate Analysis III*.
- [1981] D. R. Brillinger. Time Series Data Analysis and Theory. *Holden Day*.
- [1991] P. J. Brockwell and R. A. Davis. Time Series: Theory and Methods. *Springer, New York*.
- [1996] P. J. Brockwell and R. A. Davis. Introduction to Time Series and Forecasting. *Springer, New York*.
- [1997] A. M. Bruckner, J. B. Bruckner and B. S. Thomson. Real Analysis. *Prentice Hall*.
- [1966] R. B. Cattell. The scree test for the number of factors. *Mult. Behav. Res.* 1, 245-276.
- [1966] R. B. Cattell and J. Jaspers. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research *Mult. Behav. Res. Monographs*, 67-3, 1-212.
- [1998] M. J. Chambers. Long Memory and Aggregation in Macroeconomic Time Series. *International Economic Review*. 39, 4.
- [1994] K. S. Chan and H. Tong. A note on noisy chaos. *Journal of the Royal Statistical Society. Series B (Methodological)*. 301-311.

- [2001] K. S. Chan and H. Tong. Chaos: a statistical perspective. *Springer*.
- [2006] R. G. Clegg. A practical guide to measuring the Hurst parameter. *Arxiv preprint math.ST/0610756*.
- [1994] N. Crato and P. J. F. De Lima. Long-range dependence in the conditional variance of stock returns. *Economics Letters* 45, 281-285.
- [1996] N. Crato and K. R. Ray. Model Selection and Forecasting for Long-Range Dependent Processes. *Journal of Forecasting*. 15, 107-125.
- [1989] R. Dahlhaus. Efficient Parameter Estimation for Self-Similar Processes. *The Annals of Statistics* 17,4, 1749-1766.
- [1987] R. B. Davis and D. S. Harte. Tests for Hurst effect. *Biometrika* 74, 95-102.
- [1994] S. Dawson, C. Grebogi, T. Sauer, and J. A. Yorke. Obstructions to Shadowing When a Lyapunov Exponent Fluctuates about Zero. *Phys. Rev. Lett.* 73.
- [1986] R. L. Devaney. An Introduction to Chaotic Dynamical Systems, *Addison-Wesley, Menlo Park California*, 2.
- [1990] P. Diggle. Time Series: A Biostatistical Introduction. *Oxford university Press*.
- [1960] J. Durbin. The fitting of time series models. *Rev. Inst. Int. Stat* 28, 233-243.
- [1969] G. Durbin. Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika* 1(56).
- [1999] J. Faraway, D. B. Chaffin and X. Zhang. Simulating Reach Motions. *Society of Automotive Engineers*.
- [2007] J. Faraway and M. P. Reed. Statistics for Digital Human Motion modelling in Ergonomics.
- [2007] J. Faraway, M. P. Reed and J. Wang. Modelling three-dimensional trajectories by using B—zier curves with application to hand motion. *Appl. Statist.* 56, 5, 571-58.
- [1971] S. A. Farmer. An investigation into the results of principal component analysis of data derived from random numbers. *Statistician*, 20, 63-72.
- [1986] M. T. Fisher, J. Lee and M. K. Mara. Techniques for evaluating control of automated multi-determinant analytical instruments by computer. *Analyst* 111. 1225-1229.

- [1918] R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society* 52, 399-433.
- [1925] R. A. Fisher. Statistical Methods for Research Workers. *Oliver and Boyd, Edinburgh*.
- [1935] R. A. Fisher. The Design of Experiments. *Oliver and Boyd, Edinburgh*.
- [1929] R. A. Fisher. Tests of Significance in Harmonic Analysis. *Proceedings of the Royal Society* 125, 54-59.
- [1986] R. Fox and M. S. Taquq. Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics* 14, 517-532.
- [1983] J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series. *Journal of Time Series Analysis* 4, 221-238.
- [1990] L. Giraitis and D. Surgailis. A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probability Theory and Related Fields* 86, 87-104.
- [1999] L. Giraitis and M. S. Taquq. Whittle Estimator of Finite-Variance Non-Gaussian Time Series With Long Memory. *Annals of Statistics* 27, 1, 178-203.
- [1988] L. Glass and M. C. Mackey. From Clocks to Chaos. *Princeton University Press*.
- [2004a] S. Gouëzel. Vitesse de décroissance et théorèmes limites pour les applications non uniformément dilatantes. *PhD Thesis, Ecole Normale Supérieure*.
- [2004b] S. Gouëzel. Sharp polynomial estimates for the decay of correlations. *Israel J. Math.* 139, 29-65.
- [1980] C. W. J. Granger. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14, 227-238.
- [1981] C. W. J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121-130.
- [1980] C. W. J. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15-39.
- [1986] H. L. Gray and W. A. Woodward. A new ARMA spectral estimator. *Journal of the American Statistical Association* 81, 396, 1100-1108.

-
- [1990] C. Grebogi, S. M. Hammel, J. A. Yorke and T. Sauer. Shadowing of Physical Trajectories in Chaotic Dynamics: Containment and Refinement. *Phys. Rev. Lett.* 65, 1527-30.
- [2005] D. Guegan. How can we Define the Concept of Long Memory? An Econometric Survey. *Economic reviews*.
- [1954] L. Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*. 19, 149-161.
- [1898] J. Hadamard, Les surfaces à courbures opposés et leurs lignes géodésiques. *J. Math. Pures et Appl.* 4 , 27-73.
- [1996] P. Hall. Defining and measuring long-range dependence. *non-linear Dynamics and Time Series*.
- [1980] E. J. Hannan. The estimation of the order of an ARMA process. *Ann. Statist.* 8, 1071-81.
- [1979] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *J. R. Statist.Soc. B* 41, 190-5.
- [1982] E. J. Hannan and J. Rissanen. Recursive estimation of mixed autoregressive moving average order. *Biometrika* 69, 81-94.
- [1976] H. H. Harmon. Modern Factor Analysis. *University of Chicago Press* 3 rd Ed, Chicago.
- [1989] J. Haslett and A. E. Raftery. Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource (with Discussion). *Applied Statistics* 38, 1-50.
- [2005] W. Hayes and K. R. Jackson. A survey of shadowing methods for numerical solutions of ordinary differential equations. *Applied Numerical Mathematics* 53,2, 299-321.
- [1997] C. C. Heyde and Y. Yang. On Defining Long-Range Dependence *Journal of Applied Probability*, 34, 4, 939-944.
- [2000] R. C. Hilborn. Chaos and non-linear dynamics: An introduction for scientists and engineers, *Oxford University Press*.
- [1971] D. V. Hinkley, . Inference about the change-point from cumulative sum tests, *Biometrika*, 58 3, 509-523.

-
- [1987] D. V. Hinkley and E. Schechtman, Conditional bootstrap methods in the mean-shift model, *Biometrika*, 74 1, 85-93.
- [2005] M. P. Holland. Slowly mixing systems and intermittency maps. *Ergodic Theory and Dynamical Systems*. Volume 25, 133-159.
- [1981] J. R. M. Hosking. Fractional differencing. *Biometrika*. 68, 1, 165 - 176.
- [1933] H. Hotelling. Analysis of a Complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417-441, 498-520.
- [2000] S. Hwang. The Effects of Systematic Sampling and Temporal Aggregation on Discrete Time Long Memory Processes and Their Finite Sample Properties. *Economic Theory*. 16, 347-372.
- [1967] P. J. Huber. The behaviour of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*. 1, 221-223.
- [1951] H. E. Hurst. Long-Term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*. 116, 770-799.
- [1957] H. E. Hurst. A suggested statistical model of some time series that occur in nature. *Nature*. 180, 494.
- [1985] C. M. Hurvich. Data-driven choice of a spectrum estimate: extending the applicability of cross-validation methods. *J. Am. Statist. Assoc.* 80, 933-40.
- [1993] C. M. Hurvich and K. I. Beltrao. Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series. *Journal of Time Series Analysis*. 14, 455-472.
- [1994] C. M. Hurvich and K. I. Beltrao. Automatic Semiparametric estimation of the memory parameter of a long-memory time series. *Journal of Time Series Analysis*. 15, 3, 285-302.
- [2001] C. M. Hurvich and J. Brodsky. Broadband Semiparametric Estimation of the Memory Parameter of a Long-Memory Time Series Using Fractional Exponential Models. *Journal of Time Series Analysis*. 22, 2, 221-249.
- [1998] C. M. Hurvich R. Deo and J. Brodsky. The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis*. 19, 1, 19-46.
- [1995] C. M. Hurvich and B. K. Ray. Estimation of the memory parameter for non-stationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis*. 16, 1, 17-41.

-
- [1997] A. Inoue. Regularly varying correlation functions and KMO-Langevin equations. *Hokkaido Mathematical Journal*. 2, 26, 457-482.
- [2000] A. Inoue. Asymptotics for the Partial Autocorrelation Function of a Stationary Process. *J. Anal. Math* 81, 65-109.
- [1991] J. E. Jackson. A user's guide to Principal Components. *Wiley Series in Prob. and Math. Stat.*
- [1982] G. J. Janacek. Determining the degree of differencing for time series via the log spectrum. *J. Time Ser. Anal.* 3, 177-83.
- [1970] I. T. Jolliffe. Redundant Variables in Multivariate Analysis. *Unpublished D.Phil thesis* University of Sussex.
- [1972] I. T. Jolliffe. Discarding variables in Principal Component Analysis. I: Artificial data. *Appli. Stat.* 21, 160-173
- [1986] I. T. Jolliffe. Principal Component Analysis. *Springer-Verlag New York* 1
- [2002] I. T. Jolliffe. Principal Component Analysis. *Springer-Verlag New York, Incorporated* 2
- [1975] R. H. Jones. Fitting autoregressions. *Journal of American Statistics Association*. 70, 590-592.
- [1960] H. F. Kaiser. The application of electronic computers to factor analysis. *Educ. and Psych. Meas.* 20, 141-151.
- [1995] P. S. Kokoszka and M. S. Taqqu. Fractional ARIMA with stable innovations. *Stochastic Processes and Their Applications*. 60, 19-47.
- [1987] H. R. Künsch. Statistical aspects of self-similar processes. *Proceedings of the First World Congress of the Bernoulli Society* (Y. Prohorov and V. V. Sazanov, eds.) 1, 67-74.
- [1981] W. Lauterborn. Subharmonic Route to Chaos in Acoustics. *Phys. Rev. Lett.* 47, 1445.
- [1991] A. J. Lawrance. Directionality and reversibility in time series. *International Statistical Review/Revue Internationale de Statistique* 67-79
- [1992] A. J. Lawrance. Uniformly distributed first-order autoregressive time series models and multiplicative congruential random number generators. *Journal of Applied Probability* 896-903

- [2001] A. J. Lawrance. Statistical aspects of chaos, a series of 7 lectures for MSM4S2. *Technical Report, School of Mathematics and Statistics, University of Birmingham.*
- [2001] A. J. Lawrance and N. Balakrishna. Statistical aspects of chaotic maps with negative dependence in a communications setting. *Journal of the Royal Statistical Society. Series B, Statistical Methodology.* 843-853.
- [2005] A. J. Lawrance and R. M. Iilliam. Chaos communication synchronization: Combatting noise by distribution transformation. *Statistics and Computing* 15, 1, 43 - 52
- [1998] A. J. Lawrance and N. M. Spencer. Curved chaotic map time series models and their stochastic reversals. *Scandinavian Journal of Statistics* 371-382
- [1995] A. J. Lawrance, A. Tsuneda and T. Kohda. On higher-order correlations of chaotic sequences. *Technical Report Inst. Electronics, Informations and Communication Engineers of Japan.*
- [2003] A. J. Lawrance and R. C. Wolf. Binary Time Series Generated by Chaotic Logistic Maps. *Stochastics and Dynamics* 3, 4, 529-544
- [1993] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *D. P. Sidhu, editor, Proc.* 183-193.
- [1991] W. E. Leland and D. V. Wilson. High time-resolution measurement and analysis of lan traffic: Implications for lan interconnection. *Proc. IEEE INFOCOM* 1360-1366.
- [1947] N. Levinson. The Wiener RMS error criterion in filter design and prediction, *J. Math. Phys.* 25,261-278.
- [1975] T. Y. Li and J. A. Yorke. Period Three Implies Chaos. *Am. Math. Monthly.* 82, 985.
- [1999] C. Liverani, B. Saussol and S. Vaienti. A probabilistic approach to intermittency. *Ergodic Theory and Dynamical Systems.* 19(3): 671-685.
- [1963] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.* 20,1 , 130-141.
- [1972] E. N. Lorenz. *Talk: Predictability: Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas. American Association for the Advancement of Science.*
- [1991] H. Lutkepohl. Introduction to Multiple Time Series Analysis. *Springer-Verlag, Berlin.*

- [2001] H. Lutkepohl. Vector autoregressions. *A Companion to Theoretical Econometrics* 678-699.
- [1997] H. Lutkepohl and J. Breitung. Impulse response analysis of vector autoregressive processes *System Dynamics in Economic and Financial Models*, Wiley.
- [2004] H. Lutkepohl and M. Kratzig. Applied time series econometrics *Cambridge University Press*.
- [1968] B. B. Mandelbrot and J. Wallis. Noah, Joseph and operational hydrology. *Water Resources Research*. 4, 909-918.
- [1972] B. B. Mandelbrot. Statistical methodology for non periodic cycles: From the covariance to R/S analysis. *Annals of Economic and Social Measurement*. 1, 259-290.
- [1980] P. Manneville and M. Pomeau. Intermittent transition to turbulence in dissipative dynamical systems. *Comm. Math. Phys.* 74, 189-197.
- [1943] H. B. Mann and A. Wald. On the statistical treatment of linear stochastic difference equations. *Econometrica*. 11, 173-200.
- [1992] R. J. Martin and J. A. Eccleston. A new model for slowly-decaying correlations. *Statist. Probab. Lett.* 13, 139-145.
- [1997] R. J. Martin and A. M. Walker. A power-law model and other models for long-range dependence. *J. Appl. Probab.* 34, 675-679.
- [1978] A. I. McLeod and K. W. Hipel. Preservation of the rescaled adjusted range, 1: A reassessment of the Hurst phenomenon. *Water Resources Research*. 14, 491-508.
- [2000] A. Menache. Understanding Motion Capture for Computer Animation and Video Games. *Morgan Kaufmann*.
- [1999] R. J. Mondragon. A model of packet traffic using a random walk model. *Int. J. Bifur. Chaos*. 7, 1381-1392.
- [1979] F. C. Moon and P. J. Holmes. A Magnetoelastic Strange Attractor. *J. Sound Vib.* 65, 285.
- [1990] J. Mork, J. Mark and B. Tromberg. Route to Chaos and Competition Between Relaxation Oscillations for a Semiconductor Laser with Optical Feedback. *Phys. Rev. Lett.* 65, 1999.
- [1999] E. Moulines and P. Soulier. Broadband Log-Periodogram regression of time series with Long-range Dependence. *The Annals of Statistics*. 27, 1415-1439.

- [1998] E. Moulines and P. Soulier. Data Driven Order Selection for Projection Estimator of the Spectral Density of Time Series with Long Range Dependence. *Ecole Nationale Supérieure des Telecommunications and Université d'Evry Val-d'Essonne*.
- [2002] E. Ott. Chaos in dynamical systems. *Cambridge Univ Pr*.
- [1955] E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika* 42, 523-527.
- [1957] E. S. Page. On problems in which a change in parameter occurs at an unknown point. *Biometrika* 44, 248-252.
- [2004] W. Palma and M. Zavallos. Analysis of correlation structure of square time series. *Journal of Time Series Analysis*, 25, 529-550.
- [2007] W. Palma. Long-Memory Time Series. Theory and Methods. *Wiley Series*.
- [1974] E. Parzen. Some recent advances in time series modelling. *IEEE Trans. Automatic Control*, 19, 723-729.
- [1984] J. Paulsen. Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis*, 5, 115-127.
- [1901] K. Pearson. On Lines and Planes of closest fit to systems of points in space, *Phil. Mag., Ser. B*, 2, 559-572.
- [1980] A. N. Pettitt, A simple cumulative sum type statistic for the change-point problem with zero-one observations, *Biometrika*, 67 1, 79-84.
- [2001] P. C. B. Phillips. Regression with Slowly Varying Regressors. *Yale Cowles Foundation Discussion Paper*.
- [2006] P. C. B. Phillips and K. Shimotsu. Local Whittle estimation of fractional integration and some of its variants. *Journal of Econometrics* 130, 2, 209-233.
- [1895] H. Poincaré. Analysis Situs, *Journal de l'Ecole Polytechnique ser 2*, 1, 1-123.
- [1994] D. N. Politis and J. P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* 22 (4), 2031-2050.
- [1990] S. Porter-Hudak. An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates. *Journal of the American Statistical Association*. 85, 338-344.
- [1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. Numerical Recipes in FORTRAN. *Cambridge University Press, Cambridge*. 85, 338-344.

-
- [1992] G. D. Quinlan and S. Tremaine. On the reliability of gravitational N-body integrations. *Royal Astronomical Society, Monthly Notices* . 259, 3.
- [1974] F. L. Ramsey. Characterization of the partial autocorrelation function. *The Annals of Statistics*. 2, 6, 1296-1301.
- [1978] J. Rissanen. modelling by shortest data description. *Automatica*. 14, 465-471.
- [1994a] P. M. Robinson. Time series with strong dependence. *Advances in Econometrics. Proceedings of the Sixth World Congress* 1: 47-95.
- [1994b] P. M. Robinson. Rates of Convergence and optimal spectral bandwidth for long range dependence. *Probability Theory and Related Fields* 99 443-473.
- [1995a] P. M. Robinson. Log-Periodogram Regression of Time Series With Long Range Dependence. *The Annals of Statistics* 23: 1048-1072.
- [1995b] P. M. Robinson. Gaussian Semiparametric Estimation of Long Range Dependence. *The Annals of Statistics* 23: 1630-1661.
- [1984] R. W. Rollins and E. R. Hunt. Intermittent Transient Chaos at Interior Crises in the Diode Resonator. *Phys. Rev. A* 29, 3327.
- [1956] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences* 42, 43-47.
- [1967] Y. A. Rozanov. Stationary Random Processes. *Holden-Day, San Francisco*.
- [1990] R. Z. Sagdeev, D. A. Usikov and G. M. Zaslavsky. non-linear Physics. *Harwood Academic Publishers, New York*.
- [1953] J. D. Sargan. An approximate treatment of the properties of the correlogram and periodogram. *J. Roy. Statist. Soc. Ser. B*, 15, 140-152.
- [2002] O. M. Sarig. Subexponential decay of correlations. *Invent. Math.* 150, 629-653.
- [1964] A. N. Sarkovskii. Coexistence of Cycles of a Continuous Map of a Line Into Itself. *Ukr. Mat. Z.* 16, 61.
- [1997] T. Sauer, C. Grebogi and J. A. Yorke. How Long Do Numerical Chaotic Solutions Remain Valid?. *Phys. Rev. Lett.* 79, 59-62.
- [1959] H. Scheffe. The Analysis of Variance. *Wiley Publications in Statistics*.
- [1995] C. M. Schmidt and R. Tcherning. The Identification of Fractional ARIMA Models *Discussion Paper of Humboldt-Universitat zu Berlin*.

- [1899] A. Schuster. The periodogram of magnetic declination. *Cambridge Phil. Soc.* 18, 18.
- [1978] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.* 6, 461-4.
- [1964] G. A. F. Seber. Orthogonality in analysis of variance, *Ann. Math. Statist.* 35, 705-710.
- [1966] G. A. F. Seber. The Linear Hypothesis: A General Theory, *Griffin's Statistical Monographs*.
- [1980] R. J. Serfling. Approximation Theorems of Mathematical Statistics. *John Wiley, New York* 18, 18.
- [1965] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika.* 52, 3-4, 591-611.
- [1968] S. S. Shapiro, M. B. Wilk and V. Chen. A Comparative Study of Various Tests for Normality. *JASA.* 63, 1343-1372.
- [1984] R. Shaw. The Dripping Faucet as a Model Chaotic System. *Aerial Press, Santa Cruz*.
- [1976] R. Shibata. Selection of the Order of an Autoregressive Model by Akaike's Information Criterion. *Biometrika.* 63 (1), 117-126.
- [2000] R. H. Shumway and D. S. Stoffer. Time Series Analysis and Its Applications. *Spronger*.
- [1963] J. L. Simonds. Application of characteristic vector analysis to photographic and optical response data. *J. Optical Soc. Amer.* 53, 968-974.
- [1982] R. H. Simoyi, A. Wolf and H. L. Swinney. One Dimensional Dynamics in a Multicomponent Chemical Reaction. *Phys. Rev. Lett* 49 245.
- [1972] Y. G. Sinai. Gibbs Measures in Ergodic Theory. *Russ. Math. Surveys* 27 (4), 21.
- [1929] E. Slutsky. Sur l'extension de la theorie de periodogrammes aux suites des quantites dependentes. *Comptes Rendues.* 189, 722-733.
- [1934] E. Slutsky. Alcuni applicazioni di coefficienti di Fourier al analizo di sequenze eventuali coerenti stazionarii. *Giorn. d. Istituto degli Atuari.* 5, 435-482.
- [2007] L. A. Smith. Chaos : a very short introduction. *Oxford University Press*.

-
- [1986] M. R. Smith, S. T. Nichols, R. M. Henkelman and M. L. Wood. Application of Autoregressive Moving Average Parametric modelling in Magnetic Resonance Image Reconstruction. *Medical Imaging, IEEE Transactions on.* 5,3,132-139.
- [1997] J. Smith, N. Taylor and S. Yadav. Comparing the bias and misspecification in ARFIMA models. *Journal of Time Series Analysis.* 18,507-527.
- [1879] G. G. Stokes. Note on searching for periodicities .*Proc. Roy. Soc.* 29, 122.
- [2003] W. Szemplinska-Stupnicka. Chaos, bifurcations and fractals around us : a brief introduction . *World Scientific.*
- [2003] M. S. Taqqu. Fractional Brownian motion and long-range dependence. *Theory and Applications of Long-Range Dependence*, 5-38.
- [1996] M. S. Taqqu and V. Teverovsky. Semi-parametric graphical estimation techniques for long memory data. *Proceedings of the Athens Conference of Applied Probability and Time Series Analysis*, New York.
- [1995] M. S. Taqqu, V. Teverovsky and W. Willinger. Estimators for long-range dependence: an empirical study. *Fractals*, 3(4):785-798.
- [1997] M. S. Taqqu and V. Teverovsky. Robustness of Whittle-type estimators for time series with long-range dependence. *Stochastic Models* . 13, 723-757.
- [1980] M. Thaler. Estimates on invariant densities of endomorphisms with indifferent fixed points. *Isr. J. Math* . 37, 303-314.
- [1967] J. W. Tukey. An introduction to the calculations of numerical spectrum analysis. *Advanced Seminar on Spectral Analysis (Ed. B. Harris)*, Wiley, New York 25 - 46.
- [1972] T. G. Vavra. Factor Analysis of perceptual change. *J. Mark. Res.* 9, 193-199.
- [1975] G. Wahba and S. Wold. Periodic splines for spectral density estimation: the use of cross validation for determining the degree of smoothing. *Commun. Statist.* 4, 125-41.
- [1931] G. Walker. On Periodicity in Series of Related Terms. *Proceedings of the Royal Society of London*, 131: 518-532.
- [1975] P. Walters. An Introduction to Ergodic Theory. *Springer-Verlag*.
- [1994] M. W. Watson. Vector autoregressions and cointegration. *Handbook of Econometrics, Vol. IV, Elsevier*.

-
- [1989] W. W. S. Wei. Time Series Analysis. Univariate and Multivariate Methods. *Addison-Wesley Publishing*.
- [1973] P. H. Wiener. The estimation of relative gas phase acidities of substituted benzoic acids from solution measurements by factor analysis. *J. Amer. Chem. Soc.*, 95, 5845-5851.
- [1954] P. Whittle. A statistical investigation of sunspot observations with special reference to H. Alven's sunspot model. *Astrophys. J.* 120, 251-260.
- [1930] N. Wiener. Generalized harmonic analysis. *Acta. Math*, 55, 117-258.
- [1953] H. Wold. A Study in the Analysis of Stationary Time Series. *Almqvist and Wiksells*, 2.
- [1968] S. Yamamoto and Y. Fujikoshi. Two-way classification designs with unequal cell frequencies. *J. Sci. Hiroshima Univ. Ser. A-1* 32, 357-370.
- [1999] L. S. Young. Recurrence times and rates of mixing. *Israel J. of Math*, 110: 153-188.
- [1927] G. U. Yule. On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London*, 226: 267-298.
- [1988] A. Zygmund. Trigonometric Series, Volume I. *Cambridge University Press, Cambridge*.