New Methods for Spectral White Noise Testing Using Wavelets with Application to Functional Time Series

Delyan Boyanov Savchev

School of Mathematics

November 2016

A Thesis Submitted to the University of Bristol in accordance with the requirements of the degree Doctor of Philosophy in the Faculty of Science

Abstract

The central topic of this thesis is the development of new white noise testing methodology based on wavelets. We propose three white noise tests based on wavelet decomposition of the raw periodogram of a univariate time series. Two of the tests are based on Haar wavelets and the third uses Daubechies' wavelets with ten vanishing moments. For the Haar-based tests we derive an approximate result for the distribution of the wavelet coefficients of the raw periodogram of a white noise process. For our third test, we derived a theoretical power function. We evaluate our tests against commonly found tests in statistical software, as well as on a range of real datasets, and find that our tests have good performance. Unlike many white noise tests, ours do not require manual tuning parameters and they are implemented in a separate `R` package `hwwntest`.

Next, we extend our univariate Haar wavelet white noise test to two-dimensional (spatial) data. We design an experiment for evaluation of its performance against systematic non-random effects in images and find that it has a good performance. Furthermore, we compare its performance to an established spatial autocorrelation test on a range of spatial autocorrelation scenarios. We also analyze a well-known spatial dataset, the Mercer and Hall (1911) wheat data, and find that our test confirms the previously found trend in the data.

Finally, we embark on application of white noise testing in functional time series. We deal with the problem of order verification of the autoregressive Hilbertian process of order one and design a procedure for that. Moreover, we refine the procedure to orders greater than one by including one of our univariate wavelet white noise tests and compare this approach with an established methodology. We find that our multistage algorithm has good performance. We also suggest an applied methodology for forecasting of autoregressive Hilbertian processes based on established theoretical results.

Author's Declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the thesis are those of the author.

SIGNED .....

Delyan Boyanov Savchev

DATE .....

Contents

Abstract i

Author's Declaration ii

1 Introduction 1

2 Literature Review — White Noise Tests and Wavelets 5

2.1 Stationary time series 5

2.2 Time Series Analysis in the Frequency Domain 6

2.2.1 The Wiener-Khintchin theorem 6

2.2.2 Periodogram analysis and white noise 7

2.2.3 Two-dimensional periodogram analysis and white noise 9

2.3 The spectral approach to white noise testing 11

2.3.1 Schuster and Fisher's white noise tests 11

2.3.2 Bartlett's test for white noise 13

2.3.3 Durbin's periodogram-based test 16

2.3.4 Other work related to Fisher's and Bartlett's tests 17

2.3.5 Directions for periodogram-based tests and this thesis 18

2.4 The autocorrelation approach to white noise testing 18

2.4.1 The Durbin-Watson test for lag one serial correlation 19

2.4.2 The Extended Durbin-Watson h-test 20

2.4.3 Box-Pierce-Ljung white noise test 21

2.4.4 Ljung-Box evaluation and lag selection 22
2.5 Contemporary work on white noise testing 22
2.5.1 Small magnitude autocorrelation 23
2.5.2 The tests from Guay et al. (2013) and Hong (1996) 24
2.5.3 Lobato-Velasco periodogram-based test 25
2.5.4 Overview of white noise tests assumptions 26
2.5.5 White noise tests summary and thesis directions 26
2.6 Wavelets and wavelet transforms 27
2.6.1 Haar wavelets in one dimension 28
2.6.2 How wavelets are constructed 29
2.6.3 The Shannon wavelet and multiresolution analysis 31
2.6.4 Multiresolution Analysis 34
2.6.5 The discrete wavelet tranform(DWT) 38
2.6.6 Vanishing moments 41
2.6.7 Haar in two dimensions 43
2.6.8 Wavelets in statistics 43
2.6.9 Spectral estimation with wavelets 46

3 Literature Review — Functional Data Analysis 51
3.1 Overview of Functional Data Analysis 51
3.1.1 An early FDA problem 52
3.1.2 Tools for Analysis 54
3.2 Functional Data Analysis Framework 55
3.2.1 Functional Space, Mean and Covariance Functions 55
3.2.2 The Ramsay and Silverman monograph. 57
3.2.3 Nonparametric FDA: Ferraty and Vieu (2006) 58
3.3 Some key tools in Functional Data Analysis 58
3.3.1 Functional Principal Components Analysis 59
3.3.2 The Need for Different Norms and Metrics 59

3.3.3 Combining of fPCA and derivative metrics 60
3.3.4 Other Aspects of Functional Data Analysis 60
3.3.5 Functional time series: Bosq(2000) 62
3.4 Functional Time Series 63
3.4.1 The Autoregressive Hilbertian Process of Order One 63
3.4.2 Notation and Theoretical Setup for ARH(1) 64
3.4.3 The Data Problem in Functional Time Series 67
3.4.4 Interesting problems in time series FDA 69

4 Univariate Wavelet White Noise Tests 71

4.1 Introduction 72
4.2 Building blocks of our tests 73
4.2.1 Basic Components 73
4.2.2 Assessing Spectral Constancy 74
4.3 A Haar wavelet test 75
4.3.1 All coefficient Haar Test 76
4.3.2 Single coefficient Haar test 77
4.4 A general wavelet test 78
4.4.1 Specification of the General Wavelet Test 80
4.4.2 Power Function of the General Test 81
4.5 Computational Details 82
4.5.1 Implementation 82
4.5.2 Empirical distribution of the wavelet coefficients 84
4.6 Univariate White Noise Simulation Study 87
4.6.1 Size Estimation for the three Wavelet Tests and Others 87
4.6.2 Power Estimation for the three Wavelet Tests and Others 90
4.6.3 Simulation Comparisons with Contemporary Papers 93
4.7 Real Data Examples 94
4.7.1 Wind Speed Example 94

4.7.2 The HADCRUT4 Global Dataset 96
4.7.3 S&P 500 Annual Log Returns 100

5 Local Alternatives and Nonlinear Models 103

5.1 Introduction 103
5.2 Local Alternatives 107
5.2.1 AR/MA(p) local alternatives scenario 107
5.2.2 Theoretical power results 108
5.2.3 Spectrum estimation and periodicities 111
5.3 Nonlinear Models 111
5.4 Simulation Study with low-magnitude parameters 115
5.4.1 What about moderate values of parameters 118
5.5 Conclusion 118

6 Two-dimensional Wavelet White Noise Tests 121

6.1 Introduction 121
6.2 Basic components for the two-dimensional test 121
6.2.1 The 2D Periodogram 122
6.2.2 Usage and properties of the two-dimensional periodogram 122
6.2.3 The Theoretical Basis 123
6.2.4 The 2D Haar Wavelet Transform 124
6.3 Distribution of the $2D$ Haar wavelet cofficients 124
6.3.1 Empirical distribution of the $2D$ wavelet coefficients 125
6.4 Two-dimensional *HWWN* test procedure 127
6.5 Spatial Statistics White Noise Tests 128
6.5.1 Popular Measures of Spatial Autocorrelation 128
6.5.2 Brief literature review of spatial autocorrelation tests 129
6.5.3 An illustrative example: the wheat data 130
6.5.4 Software for Spatial Statistics 138

6.6 Spatial Autocorrelation Testing . . . 138
6.6.1 The Matrix of Spatial Weights . . . 138
6.6.2 Weighting and the 2D Haar Wavelet Transform . . . 140
6.7 Simulation Results . . . 142
6.7.1 The Spatial Autoregressive Model (SAR) . . . 142
6.7.2 Relationship between Moran's I and SAR parameter ρ . . . 144
6.7.3 Empirical size simulations . . . 145
6.7.4 Empirical power simulations . . . 147
6.7.5 Refining the test as univariate d_{22} . . . 149
6.7.6 Homogenizing and contaminating of white noise . . . 149
6.8 Conclusion and Further Work . . . 153

7 ARH Order Verification Methodology 155

7.1 ARH(1) Order Verification . . . 155
7.1.1 Background . . . 155
7.1.2 Our methodology for verification of ARH(1) . . . 157
7.1.3 Algorithm for verification of ARH(1) — *VERARH* . . . 158
7.2 Simulation study for ARH(1) . . . 159
7.2.1 Setup of the Simulation Routine for ARH(1) . . . 159
7.2.2 Results from applying VERARH for the verification of ARH(1) . 161
7.2.3 Conclusions from the simulation Results . . . 162
7.3 Simulation Study for ARH(2) . . . 164
7.3.1 Setup of the Simulation for ARH(2) . . . 164
7.3.2 Results from applying extended VERARH for the verification of ARH(2) . . . 165
7.3.3 The upgraded VERARH procedure, using d00 wavelet test . . . 166
7.3.4 A Real Functional Time Series Example . . . 168
7.4 Simulation study with GCV and *VERARH* for ARH(1) prediction . . . 172
7.4.1 Forecasting of ARH processes using the *VERARH* method . . . 172

7.4.2 Results from Forecasting Simulated Data 177
7.5 Conclusion and Further Work 180

8 Conclusions and Innovations **181**

Bibliography **184**

A Proofs, ARH simulations and software **197**
A.1 Proof of Proposition 2 197
A.2 Proof of Proposition 3 198
A.3 Proof of Approximation 4 200
A.4 Derivation of the $d_{0,0}$ index 201
A.5 Proof of Proposition 5 202

List of Figures

2.1 Left: Scaling Function Right: Mother Wavelet Top: Haar wavelet: $m = 1$; Middle: Daubechies Extremal Phase wavelet with $m = 5$; Bottom: Daubechies Extremal Phase wavelet with $m = 10$; 42

2.2 Illustration of generic steps 1 and 2 from the two-dimensional discrete wavelet transform algorithm. $C^j := \mathbb{X}_{n\times n}$ and D^1, D^2 and D^3 are the horizontal, vertical and diagonal coefficients resepectively. Reproduced with permission from Nason (2008). 44

2.3 Realization of the AR(4) process from equation(2.70). 47

2.4 Top: The raw periodogram of the AR(4) process from equation(2.70); Bottom:Its wavelet coefficients of raw periodogram from a Daubechies extremal phase wavelet with 5 vanishing moments. 48

3.1 These curves represent the tongue dorsum height while pronouncing the sound 'Kah'. They are result of polynomial spline smoothing of tongue position sampled every milisecond using an ultrasound sensing technique. Each record begins and ends at the point where the slope is zero. Here the lengths of the curves have been standardized to the interval $(0, \pi)$. Picture reproduced with permission from Besse and Ramsay(1986) 53

3.2 EDF daily electricity load time series from September 2002 till August 2009 63

3.3 This picture shows the Electricite de France electricity load curves from 1 Sep 2008 to 31 Aug 2009. On the x axis we have the 48 instances of daily measurement and on the y is the load in kw/h 64

3.4 5 curves of a Simulated Realization of ARH(1) with Wiener Noise over a grid of 100 points 68

4.1 Top left: probability density estimate (solid) of Haar wavelet coefficients $\hat{v}_{11,k}$ with $g_2(x)$ superimposed (dotted). Top right: equivalent but for cumulative distribution. Bottom left: empirical histogram of p-values. . . . 77

4.2 All plots: solid line is theoretical power calculated from (4.12) for the general wavelet test, circles are results from simulation study below. Top left: independent and identically distributed normal random variables; Top right: AR(1) with $\alpha = 0.3$, Table 4.2; Bottom left: MA(2) with $\beta_1 = 0, \beta_2 = 0.5$, Table 4.3, Bottom right: AR(12) with $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$, Table 4.3. 83

4.3 Top left to bottom right respectively: black — Empirical distribution of finest-scale Haar wavelet coefficients of 10^3 realizations from Gassian white noise with $T = 128, 256, 512, 1024$, red — theoretical Macdonald curve 86

4.4 Top left to bottom right respectively: black — Empirical distribution of 10^3 realizations from Gassian white noise with $T = 32, 64, 128, 256$, red — standard Gaussian curve $N(0, 1)$ 88

4.5 Shapes of the Macdonald distribution with varying the m parameter; red — standard Gaussian curve $N(0, 1)$ 89

4.6 Top left: Aberporth wind speed time series. Top right: first differences of Aberporth wind speed time series. Bottom left: Autocorrelation function of Aberporth wind speed first differences. Bottom right: Cumulative normalized periodogram (solid black) and ideal white noise line (red dotted) for Bartlett white noise test. 95

4.7 Hadcrut4: Global ensemble medians of temperature anomalies from 01/1850 till 04/2014 96
4.8 Hadcrut4: Global ensemble medians of temperature anomalies from $01/1850$ till $04/1935$ (first 1025 observations of the data from Fig. 4.7 97
4.9 ACF and PACF of the temperature anomalies monthly raw data from $01/1850$ till $04/2014$. 97
4.10 ACF and PACF of the temperature anomalies monthly raw data from $01/1850$ till $04/1935$. 97
4.11 ACF and PACF of first order diffences of the temperature data. 98
4.12 Auto and Partial (auto) correlations of the residual series from ARIMA(1,1,1) model on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset 99
4.13 Annual Log-Returns from S&P 500 from 1871 till 1998 101
4.14 ACF and PACF of the lLog-Returns from S&P 500 from 1871 till 1998 . 102

5.1 Top left: Spectrum of AR(3) process with parameter $\alpha_i = 0.1767$ for $i = 1, 2, 3$ corresponding to local alternatives scenario $\alpha = 2/\sqrt{(T)}, \quad T = 128$. Top right: the Haar wavelet coefficients of the normalised spectrum with $T = 128$. Bottom left and right: spectrum of AR(3) with negative parameters and its Haar wavelet coefficients for $T = 128$ 104
5.2 Top left: Spectrum of AR(1) process with parameter $\alpha_i = -0.6$. Top right: the Haar wavelet coefficients of the normalised spectrum with $T = 128$. Bottom left and right: spectrum of AR(1) with parameter $\alpha_i = -0.9$ and its Haar wavelet coefficients for $T = 128$ 106
5.3 Spectra and their Haar wavelet coefficients for local alternatives Scenario 1a. Top: AR(1), middle: AR(4), bottom: AR(6). 112
5.4 Spectra and their wavelet coefficients (10 vanishing moments) for local alternatives Scenario 1a. Top: AR(1), middle: AR(4), bottom: AR(6). . . 113
5.5 AR1 ($\rho_1 = 0.1$) Power 117

5.6 AR1 ($\rho_1 = 0.125$) Power 117

5.7 AR1 ($\rho_1 = 0.15$) Power 118

5.8 MA1 ($\theta_1 = 0.1$) Power 119

5.9 MA1 ($\theta_1 = 0.15$) Power 119

5.10 AR1 ($\rho_1 = 0.3$) Power 120

6.1 Two-dimensional 128×128 white noise as a grey-scale image 122

6.2 Top left to bottom right respectively: black — 1000 realizations of Empirical distribution of finest-scale diagonal $2D$ Haar wavelet coefficients of 10^3 realizations from Gaussian white noise with $T = 256, 1024, 4096, 16384$, red — theoretical Macdonald curve 126

6.3 Wheat Raw Data Total Surface 132

6.4 Wheat Data — Residuals Surface after Median-Polish 133

6.5 256 Wheat datapoints subset's surface — raw data 134

6.6 256 Wheat datapoints subset's surface — residuals after *median-polish* . . 134

6.7 Wheat 256 datapoints subset's images — total (left) and residuals after median polish(right) 135

6.8 Image of 256 observations of Gaussian white noise 135

6.9 Total column effects vs plot number (east-west) (left) and the same for our 256-datapoint subset(right) 136

6.10 Left: image of the original 20×25 wheat data. Right: image of the augmented 32×32 wheat data. 137

6.11 Left: 2D periodogram of 32×32 Gaussian white noise. Right: 2D periodogram of 32×32 SAR model. 141

6.12 Left: 2D periodogram of 32×32 Gaussian white noise. Right: 2D periodogram of 32×32 SAR model. 141

6.13 Left: image of the 2D periodogram of 32×32 Gaussian white noise. Right: image of the 2D periodogram of 32×32 SAR model ($\rho = -0.8$). . 142

6.14 The empirical distribution of Moran's I for a SAR model with $\rho = 0.5$ for 1000 realizations from a 32×32 grid 144

6.15 Histogram of ordinary p-values from 2D *HWWN* test of a 128×128 white noise data — for all coefficients at all scales 146

6.16 Image of 128×128 observations of Gaussian white noise with mean of 3 and variance 4 152

6.17 Image of 128×128 Gaussian white noise, after the *homogenizing* operation 152

6.18 Image of 128×128 Gaussian white noise, after the *contaminating* operation 152

7.1 left: distribution of the value of the autocorrelation coefficient, defined as $\alpha_j = 0.8$ for the first 50 pricinpal components and right: its p-value, for all 1000 realizations with $n = 500$ curves and p=100 discretizations points/principal components 163

7.2 left: distribution of the value of the autocorrelation coefficient, defined as $\alpha_j = 0.5$ for the first 50 pricinpal components and right: its p-value, for all 1000 realizations with $n = 500$ curves and p=100 discretizations points/principal components 163

7.3 Percentage of explained variance by the first 5 principal components for 1000 realizations of ARH(2) 166

7.4 EDF daily electricity load time series from September 2002 till August 2009 169

7.5 All the curves for the last 1 year of data i.e. Sep 2008 - Aug 2009 170

7.6 Autocorrelations for transformed daily data taken at the: a - 1st gridpoint(00:30 hrs), c - 24th gridpoint(12:00 hrs), e - 36th gridpoint(18:00 hrs)

Partial autocorrelations for transformed daily data taken at the: b - 1st gridpoint(00:30 hrs), d - 24th gridpoint(12:00 hrs), f - 36th gridpoint(18:00 hrs) 171

7.7 Plot of typical MAPE distribution over the 100 curves forecast horizon for the $\widetilde{\boldsymbol{\rho_1}}$ predictor. Green — our method for choosing k, Red — generalised CV 179

7.8 Plot of typical MAPE distribution over the 100 curves forecast horizon for the $\widetilde{\boldsymbol{\rho_2}}$ predictor. Green — our method for choosing k, Red — generalised CV 179

List of Tables

4.1 Empirical size for the five white noise tests for various sample sizes, T. True model is independent and identically distributed variates. Approximate theoretical power from (4.12) computed to be 4.9% for all T for the General Wavelet Test genwwn. 90

4.2 Empirical power for the five white noise tests for various sample sizes, T. True model is AR(1) with parameter α with standard normal innovations. 91

4.3 Empirical power for the five white noise tests for various sample sizes, T, all with standard normal innovations. MA(2) model has $\beta_1 = 0, \beta_2 = 0.5$, ARMA$(1, 2)$ model has $\alpha_1 = -0.4$, $\beta_1 = -0.8, \beta_2 = 0.4$, AR(12) model has $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$. 92

4.4 Empirical power for the three white noise tests for $T = 256$ for different models with Student's t distributed noise with two degrees of freedom in roman font. Results with Gaussian innovations, reproduced from tables above, are in italic font. Model: a.) AR(1) $\alpha = 0.3$ from Table 4.2; b.) MA(2), $(\beta_1 = 0, \beta_2 = 0.5)$ from Table 4.3; c.) AR(12), $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$ from Table 4.3. 92

4.5 Results from applying white noise tests to the residuals of ARIMA(1,1,1) on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset 100

4.6 Results from applying the Ljung-Box test with different lags to the residuals of ARIMA(1,1,1) on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset 100
4.7 Results from applying white noise tests to the log-returns from S&P 500 for 1871-1998 101
4.8 Results from applying the Ljung-Box test with different lags to the log-returns from S&P 500 for 1871-1998 101

5.1 Approximate theoretical power for d_{00} in Scenario 1: AR(p) 109
5.2 Approximate theoretical power for d_{00} in Scenario 1: MA(q) 109
5.3 Approximate theoretical power for *hwwn* in Scenario 1: AR(p) 109
5.4 Approximate theoretical power for d_{11} in Scenario 1: AR(p) 109
5.5 Approximate theoretical power results for local Scenario 1a: AR(p) . . . 110
5.6 Approximate theoretical power results for local Scenario 2a: MA(q) . . . 110
5.7 Approximate theoretical power results for local Scenario 1a: AR(p) . . . 111
5.8 Approximate theoretical power results for local Scenario 1a: MA(q) . . . 111
5.9 Empirical power of various tests on 10^3 realizations following *tmWN* with $T = 1024$ 114
5.10 Empirical power results for various white noise tests against GARCH$(1, 1)$ models m_1 to m_6: 10^3 relizations with $T = 1024$ 115

6.1 Statistical Size for $MVN \sim (\mathbf{0}, I)$ for Moran's test, our HWWN and HWWN(BH) - with using False Discovery Rate instead of Bonferroni. . . 145
6.2 Statistical Size for $U \sim (0, 1)$ 146
6.3 Statistical Power for Gaussian UAR with $\rho = 0.3$ 147
6.4 Statistical Power for Gaussian UAR with $\rho = 0.5$ 147
6.5 Statistical Power for Gaussian UAR with $\rho = 0.7$ 147
6.6 Statistical Power for Gaussian SAR with $\rho = 0.2$ 148
6.7 Statistical Power for Gaussian SAR with $\rho = 0.5$ 148

6.8 Statistical Power for Gaussian SAR with $\rho = 0.8$ 149

6.9 Statistical Power for Gaussian SAR with $\rho = -0.3$ 149

6.10 Statistical Power for Gaussian SAR with $\rho = -0.5$ 149

7.1 Empirical power (left) and size (right), testing 100 principal components for 1000 realizations of ARH(1) with *VERARH* 161

7.2 Results from Ljung-Box test up to lag 6, testing the residuals of AR(2) fits to the first 5 principal components for 1000 realizations of ARH(2) . . 165

7.3 Results from standard PCA and the method for the transformed EDF data. 170

7.4 Table of the forecast comparisons using our method versus the CV method for defining the number of eigenvalues to retain in the expansion of the lag 0 covariance for ρ_1 and symmetrized lag 1 autocorrelation for ρ_2 respectively 178

Chapter 1

Introduction

Time series analysis is a branch of statistics that deals with data observed on a time grid, often at regular intervals, which are assumed to be a realization of a discrete stochastic process $\{X_t\}_{t=1}^{T}, T \in \mathbb{N}$. However, the intervals might also be irregular. The main difference with other branches of statistics is that, because we are observing the same process or random variable at different instances through time, dependence structure is present in the data. For some time series, autocorrelation can be found among the regular observations, but for others such as financial time series, it exists between the squared values of the observations and can be modelled by GARCH models as in Bollerslev (1986). For successful modelling of time series, this correlation structure must be taken account of. Thus, there are different approaches to time series analysis. For example, we might consider estimation of the correlation in the time domain or its dual, the spectrum, in the frequency domain. Moreover, since we expect correlation structure in the observed data, it is possible that the data-generating process might be decomposed in different mathematical bases. Furthermore, we need to make assumptions about the character of the dependence in order to choose the right mathematical basis. This ultimately depends on the objective of the statistical modelling of the time series. For instance, prediction of future values of the process e.g. inflation or stock prices indices, electricity consumption or internet traffic. Another form of application might be to classify time-dependent

patterns of curves e.g. in meteorology we have different weather stations and record measurements at fixed instances throughout the day. Then, based on those classifications, we could derive physical and/or statistical models for doing weather forecasts for the different regions of interest.

The concept of white noise testing is central to time series model validation and testing for (auto)correlation structure. This is because when we believe that a dataset follows a specific model, we should validate that in a certain way. As in simple regression where we want the residuals to be independent and identically distributed as Gaussian or close to, in the same way in time series analysis, such a requirement is to have no autocorrelation structure in the residuals from a model. That condition is often crucial for the limit theorems to hold so that we are able to derive meaningful confidence intervals, for instance.

This thesis has two main purposes:

1. The development of new methods for univariate and spatial white noise testing via wavelet analysis of the raw periodogram.

2. The analysis, via verification of the order, and forecasting of, functional time series, specifically the Autoregressive Hilbertian Processes (ARH).

Chapter 2 reviews some time and frequency approaches for white noise testing and related tests. We start with a key result from Brockwell and Davis (1991) that derives the distribution of the raw periodogram of white noise, which is central to our further work. While white noise testing itself has played a historical part in the development of random number generators and has other applications in engineering, we do not cover those. We focus on the development of white noise tests within time series analysis during the previous century in both the time and frequency domain, which has naturally evolved together with statistical science. We try to emphasize the communalities and different features of

the various tests in the review. Moreover, we discuss how are they related to each other and how that influences their performance. Further, chapter 2 provides a brief review of wavelets and multiresolution analysis and then elaborates on aspects of wavelet shrinkage that relate to our proposed work of white noise testing.

Chapter 3 reviews the relatively new branch of statistics called functional data analysis. This is a branch of statistics that deals with modelling of curves and surfaces. Functional data arise when the object of interest is a mathematical function. Typically, though, we are only able to obtain samples of the function, often contaminated with noise and/or blurred. In this thesis we examine functional time series which arise as we consider two time scales operating on the same functional object e.g. observe every day Earth's magnetic field every 100 miliseconds during the day. In this review we focus on the basic principles for performing functional data analysis drawing on monographs in the literature. We also discuss a basic model for functional time series called the autoregressive Hilbertian process. This is a very popular and frequently applied model in the field and subject to our subsequent investigation.

Chapter 4 develops three univariate white noise tests based on wavelet analysis of the raw periodogram of a time series. The first test uses Haar wavelets and we derive an approximate result for the exact distribution of the wavelet coefficients. Then we suggest a multiple comparisons procedure to test all wavelet coefficients of the raw periodogram in order to judge whether or not our series are white noise. The second test is based on a single coarsest-scale wavelet coefficient and thus has a simple form. Moreover, it turns out that it can be expressed as a weighted sum of autocorrelations of the time series. Our third wavelet test uses Daubechies' wavelets with ten vanishing moments and we derive a normal approximation based on a result from Neumann (1996). We developed a theoretical power function for this test against an alternative hypothesis of ARMA class, which can give guidance on the sample size needed in order to attain certain power level. We

conclude the chapter with an extensive simulation study considering also non-Gaussian models and comparing our tests against other tests found in statistical software. We also present three different real datasets which benefit from white noise analysis.

Chapter 5 demonstrates that white noise tests have application in time series model order verification for univariate AR and MA models. We focus on the case of moderate number of observations $(100 - 500)$ and parameters of small magnitude. We performed a simulation study which showed that usual AIC usage delivers 75% statistical power, whereas by using white noise tests we can get the power to 100% for our study.

Our two-dimensional extension of the Haar wavelet test is proposed in chapter 6. We investigate the hypothesis of spatial autocorrelation and compare our test with one of the standards in the literature. Furthermore, we analyze one of the classical spatial datasets: the Wheat data from Mercer and Hall(1911) and show that our test correctly detects the trend explored by many authors before. We also design a simulation experiment with non-random operations on an image and explore how our test deals with those. We also give brief guidance to further research and discuss how our work could be improved.

Chapter 7 suggests a multistage procedure for verification of the order of the autoregressive Hilbertian process. First, we explore order one processes and then extend the procedure to larger orders. White noise testing is central to our algorithm and we compare its results with an established method for the order verification problem for ARH. We show a practical example with electricity load data. Based on our routine, we also suggest a procedure for forecasting and compare it with a well-known generalised cross-validation approach from the literature on simulated data. We find that our routine has a good performance.

Finally, we conclude in chapter 8 and discuss possible further research.

Chapter 2

Literature Review — White Noise Tests and Wavelets

2.1 Stationary time series

A key concept in time series analysis is *stationarity*. Colloquially speaking, stationarity means that the statistical properties of the stochastic process do not depend on the absolute value of time. However, when defining it mathematically there are different options. At the highest level of taxonomy in the literature, there exist two main forms of stationarity - strict stationarity and wide-sense stationarity with the latter being most popular. Let $X_t, t \in \mathbb{N}$ is a time series. Before getting to the definitions, let us define the main statistical quantities of a time series following Chatfield (1996) chapter 3: the

1. Mean function

$$\mu(t) = \mathrm{E}\{X(t)\} \tag{2.1}$$

2. Variance function

$$\sigma^2(t) = \mathrm{Var}\{X(t)\} \tag{2.2}$$

3. Autocovariance function

$$\gamma(\tau) = \mathrm{E}[\{X(t) - \mu(t)\}\{X(t+\tau) - \mu(t+\tau)\}] \tag{2.3}$$

4. Autocorrelation function

$$\rho(\tau) = \gamma(\tau)/\gamma(0). \tag{2.4}$$

Here $\tau \in \mathbb{Z}$ is known as the lag.

Definition 1. *(Chatfield (1996), page 28)*

*A time series process $\{X_t\}_{t=1}^{T}$ is said to be **strictly stationary** if the joint distribution of $X_1, \ldots, X_T$ is the same as the joint distribution of $X_{1+\tau}, \ldots, X_{T+\tau}$ for all τ.*

However, this definition is usually too restrictive to employ in practice and difficult to verify. Therefore, a widely used relaxation, which is popular in practice, is a form of wide sense stationarity called **second order stationarity**.

Definition 2. *(Chatfield (1996), page 29)*

*A times series process $X(t)$ is said to be **second order stationary** if $\mathrm{E}\{X(t)\} = \mu$, $\mathrm{Cov}\{X(t), X(t+\tau)\} = \gamma(\tau)$ and $\mathrm{E}(X_t^2) < \infty$.*

From now on, we will assume that $X_t, t \in \mathbb{Z}$ is a second order stationary process, unless otherwise stated.

2.2 Time Series Analysis in the Frequency Domain

In this section, we will present the main results upon on which we will later develop our white noise tests in chapters 4 and 6.

2.2.1 The Wiener-Khintchin theorem

The Wiener-Khintchin theorem is a key result on which the theory and practice of second order stationary stochastic processes is built.

Theorem 1. *(Wiener-Knintchin, Priestley (1983), page 219) A necessary and sufficient condition for $\rho(\tau)$ to be the autocorrelation function of some stochastically continuous stationary process, $\{X_t\}$, is that there exists a function $F(\omega)$, having the properties of a distribution function on $(-\infty, \infty)$, (i.e. $F(-\infty) = 0$, $F(+\infty) = 1$, and $F(\omega)$ non-decreasing), such that for all τ, $\rho(\tau)$ may be expressed i the form:*

$$\rho(\tau) = \int_{-\infty}^{+\infty} \exp\{i\omega\tau\} dF(\omega), \tag{2.5}$$

where the integral is a Stieltjes integral.

Moreover, if the derivative $f(\omega) = dF(\omega)/d\omega$ exists, the theorem says that the autocorrelation function is the Fourier transform of the normalized power spectral density of the stochastic process X_t. The spectral density is called normalized since $\rho(\tau) = \gamma(\tau)/\gamma(0)$, thus a similar relation holds for the autocovariance function $\gamma(\tau)$. Therefore, the autocovariance and the power spectral density are Fourier pairs. Hence, the inverse relation to equation (2.5), which defines the normalized spectrum is (Priestley (1983), page 216):

$$f(\omega) = (2\pi)^{-1} \int_{-\infty}^{+\infty} \exp(-i\omega\tau)\rho(\tau) d\tau, \tag{2.6}$$

In other words, the power spectral density describes the distribution of the variance over frequency rather than time. This is also very useful in practice, since it would be hard to judge from a graph what is the main periodic component of a time series. In many applications we collect data with predefined sampling rate such as daily, weekly, monthly, quarterly or annually, thus different effects and seasonalities can be present in the observed data.

2.2.2 Periodogram analysis and white noise

A key estimator of the spectrum of a second-order stationary time series process is the periodogram from Brockwell and Davis (1991), page 342 which can be estimated from a

realization $\{X_t\}_{t=1}^{T}$:

$$I_T(\omega) = (2\pi T)^{-1} \left| \sum_{t=1}^{T} X_t e^{-i\omega t} \right|^2, \tag{2.7}$$

which can be computed at the Fourier frequencies $I_p = I_T(\omega_p)$, where $\omega_p = 2\pi p T^{-1}$ for $p = 1, \ldots, T/2$.

Definition 3. *An independent and identically distributed second-order stationary discrete stochastic process, with mean zero and variance σ^2 is denoted by $Z_t \sim IID(0, \sigma^2)$, $t = 1 \ldots T, \quad T \in \mathbb{N}, \quad \sigma^2 \in \mathbb{R}^+, \quad \sigma^2 < \infty$ is called **white noise**. If $\{Z_t\}$ are mutually independent random variables and $Z_t \sim N(0, \sigma^2)$, where N denotes the Normal (Gaussian) distribution, then it is called Gaussian white noise. Due to the properties of the Normal distribution, a sufficient condition for independence of Z_t is that $\mathrm{Cov}(Z_t, Z_{t+k}) \neq 0$, where $k \in \mathbb{Z}$ and $k \neq 0$*

A key component of the univariate white noise tests developed in this thesis, and for spectrum estimation from the periodogram in general, is the following result from Brockwell and Davis (1991), page 344, Proposition $10.3.2$, concerning the distribution of the periodogram ordinates in equation (2.7). We will note that this proposition holds for arbitrary frequencies λ_i in the range $[-\pi, \pi]$, but also for the Fourier frequencies, which are of our interest here. Next we present a modified (less general) version of this result.

Proposition 1. *Suppose that $Z_t \sim IID(0, \sigma^2)$, $\sigma^2 \in \mathbb{R}^+$, and let $I_T(\omega), -\pi \leq \omega \leq \pi$, denote the periodogram of $(Z_1, Z_2, \ldots, Z_T)$ as defined by (2.7)*

1. *If $0 < \omega_1 < \ldots < \omega_p < \pi$ then the random vector $\{I_T(\omega_1), \ldots, I_T(\omega_p)\}^T$ converges in distribution as $T \to \infty$ to a vector of independent and exponentially distributed random variables, each with mean σ^2.*

2. *If $EZ_1^4 = \eta\sigma^4 < \infty$ and $\omega_j = 2\pi j/T \in [0, \pi]$, then*

$$\mathrm{Var}\{I_T(\omega_j)\} = \begin{cases} T^{-1}(\eta - 3)\sigma^4 + 2\sigma^4, & \text{if } \omega_j = 0 \text{ or } \pi. \\ T^{-1}(\eta - 3)\sigma^4 + \sigma^4, & \text{if } 0 < \omega_j < \pi. \end{cases}$$

and

$\mathrm{Cov}\{I_T(\omega_j), I_T(\omega_k)\} = T^{-1}(\eta - 3)\sigma^4$ *if* $\omega_j \neq \omega_k$. *If* Z_1 *is Normally distributed, then* $(\eta - 3) = 0$ *so that* $I_T(\omega_j)$ *and* $I_T(\omega_k)$ *are uncorrelated for* $j \neq k$.

Proposition 1 will be a keystone on which we build our wavelet-based univariate white noise tests in chapter 4.

2.2.3 Two-dimensional periodogram analysis and white noise

Starting with a square matrix of data $\mathbb{X}_{ts}$ for $t, s = 1, \ldots, T$, $T \in \mathbb{N}$, we can calculate the two-dimensional periodogram by the following:

$$I_{T,S}(\omega_1, \omega_2) = (2\pi)^{-2} T^{-2} \left| \sum_{t=1}^{T} \sum_{s=1}^{T} X_{t,s} e^{-i(t\omega_1 + s\omega_2)} \right|^2, \tag{2.8}$$

which can be computed at the Fourier frequencies $I_{p,q} = I_{T,S}(\omega_p, \omega_q)$, where $\omega_p = 2\pi p T^{-1}$ and $\omega_q = 2\pi q T^{-1}$ for $p, q = 1, \ldots, T/2$ respectively.

Proposition 1 has its generalization in two dimensions, developed as Theorem 3.2 in Brillinger (1969). The result is that the two-dimensional periodogram has a Wishart distribution with one degree of freedom. The assumptions, on which the result relies, are strict stationarity of the series, existence of all moments and summability of the cumulants.

Assumption 1. *(Brillinger (1969), Assumption I) Let* $\mathbf{X_t}$ *is a vector-valued strictly stationary series all of whose moments exist. For each* $j = 1, 2, \ldots, k-1$ *and any* k*-tuple* $a_1, a_2, \ldots, a_k$*, with* $c_{aa}(t)$ *being the autocovariance of* $X_a(t)$ $(a = 1, 2, \ldots, s)$ *we have:*

$$\sum_{t_1, \ldots, t_{k-1}} |t_j c_{a_1, \ldots, a_{k-1}}(t_1, \ldots, t_{k-1})| < \infty \quad (k = 2, 3, \ldots) \tag{2.9}$$

In the Gaussian white noise case, because cumulants of order greater than two vanish, equation (2.9) reduces to

$$\sum_{-\infty}^{\infty} |t c_{aa}(t)| < \infty \tag{2.10}$$

Next, we present a simplified version of Theorem 3.2 from Brillinger (1969)

Theorem 2. *Let $\mathbb{X}(t)$ be a multivariate time series satisfying assumption I(Brillinger (1969)), having mean vector* **0***, and the two-dimensional periodogram $I_{T,S}$ is defined at the Fourier frequencies as in equation(2.8). Let $\mathbb{W}_S(1, V)$ denotes the Wishart distribution with one degreee of freedom and scale parameter V. If $0 \leq \omega_p \leq \pi$, then the elements of the matrix $I_{T,S}(\omega_1, \omega_2)$ are asymptotically independent. $I_{T,S}(\omega_1, \omega_2)$ tends in distribution to $\mathbb{W}_S\{1, I_{T,S}(\omega)\}$.*

Although Assumption 1 requires strict stationarity, it has been applied in practice for two-dimensional spectral estimation with second order stationarity assumptions. Examples are Pawitan (1996) and Rao et al. (2014). We will also use the result from Theorem 2 in order to infer the approximate distribution of two-dimensional Haar wavelet coefficients of the periodogram in 6. Brilliger (2001), Theorem 4.4.1 displayed next, also shows that the two-dimensional discrete Fourier transform has a multivariate complex Gaussian distribution, thus its squared magnitude would be exponentially distributed.

Theorem 3. *Let $\mathbb{X}(t)$, $t = 0, \pm 1, \ldots$ be an r vector-valued series satisfying assumption I(Brillinger (1969)), $\mathbf{c}_X = \mathrm{E}X(t)$ and . Let $s_j(T)$ be an integer with $\lambda_j(T) = 2\pi s_j(T)/T \to \lambda_j$ as $T \to \infty$ for $j = 1, \ldots, J$. Suppose $2\lambda_j(T) \pm \lambda_k(T) \not\equiv 0$(mod 2π) for $1 \leq j \leq k \leq J$. Let*

$$\mathbf{d}_X^{(T)}(\lambda) = \sum_{t=0}^{T-1} \mathbb{X}(t) \exp\{i\lambda t\} \quad -\infty < \lambda < \infty. \tag{2.11}$$

Then $\mathbf{d}_X^{(T)}(\lambda_j(T))$, $j = 1, \ldots, J$ are asymptotically independent $N_r^C(\mathbf{0}, 2\pi T \mathbf{f}_{XX}(\lambda_j))$ variates respectively. Also if $\lambda = 0, \pm 2\pi, \ldots$, $\mathbf{d}_X^T(\lambda)$ is asymptotically $N_r(T\mathbf{c}_X, 2\pi T \mathbf{f}_{XX}(\lambda))$ independently of the previous variates and if $\lambda = 0, \pm\pi, \pm 3\pi, \ldots$, $\mathbf{d}_X^T(\lambda)$ is asymptotically $N_r(\mathbf{0}, 2\pi T \mathbf{f}_{XX}(\lambda))$

Note that $\mathbf{f}_{XX}$ is the $r \times r$ spectral density matrix of the series $\mathbb{X}(t)$.

2.3 The spectral approach to white noise testing

There appear to be two main approaches for testing the null hypothesis of white noise — a time domain approach based on the sample (partial) autocorrelation function and a frequency domain approach that is based on power spectrum estimation. Historically, the spectral approach appears to have been developed first, e.g. Schuster (1898), during the end of the nineteenth and the first half of the twentieth century.

The null hypothesis, for a discrete-time second order stationary stochastic process to be a white noise, is equivalent to a flat spectral density in the frequency domain. However, in the real world, when we have a dataset to analyze, we are only dealing with one possible realization of a stochastic process. Therefore, we cannot expect a perfect flat line from the periodogram of a real dataset. However, what we can do as statisticians is to hypothesize a model for the dataset e.g. independent and identically distributed (iid) data or even iid Gaussian. Then, the spectrum of under this null hypothesis would be a flat line.

Historically, the term *periodogram* was coined in the paper of Schuster (1898), entitled "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena". The research problem boils down to the following situation: we might have a dataset from a natural phenomenon which contains obvious periodicities such as those caused by tidal semi-diurnal cycle or eleven-year sunspot maxima (Schuster (1898)). However, something such as the "lunar influence of the daily variation of magnetic forces" (Schuster (1898), page 13) might not be so easy to detect and verify. Schuster (1898) proposed to look at the problem by considering the "probability for different values of the amplitudes if the original numbers are chosen at random"

2.3.1 Schuster and Fisher's white noise tests

This section follows closely Fisher (1929). The starting assumptions of Schuster's test test is that we have a realization $X_1, X_2, \ldots, X_{2t+1}$ and $X_t \sim N(0, \sigma^2), \sigma_2 \in \mathbb{R}^+$. Thus

any linear combination of Xs would be also normally distributed and for orthogonal linear combinations the independence will also hold. Thus Schuster (1898) arrived at the main corollary from our Proposition 1. If $A = \sum_{r=1}^{2n+1} a_r X_r$ and $B = \sum_{r=1}^{2n+1} b_r X_r$, then $z = A^2 + B^2$ will be exponentially distributed i.e. $z \sim Exp(2\sigma^2)$ so that the probability of exceeding a given value of z is $\exp\{-z/2\sigma^2\}$. The test consisted of testing every particular value. The drawback is that the population variance σ^2 is not known in advance.

Section 2 of Fisher (1929) acknowledges further work on the topic by Sir Gilbert Walker, but an explicit paper is not cited however. Walker's idea is to test the significance of the largest periodogram ordinate in relation to its other values — considering the ratio of the largest periodogram ordinate to the sum of all. If $P = \exp^{-z/2\sigma^2}$ is a small probability that is to be used as a significance level, then $(1 - \exp\{-z/2\sigma^2\})^n = 1 - P$ is the relation which has to be solved in order to get an empirical critical value for z.

Fisher (1929) extends the test of Walker to test any value of the periodogram, not necessarily the largest, but second or less largest. He also considers the resulting sampling error from the unknown variance σ^2 and shows that, if we use ratio of periodogram values (i.e. ratio of exponential random variables), then it will be F-distributed, thus σ^2 is irrelevant, since it is replaced by the usual sum-of-squares sample variance formula i.e. using $T - 1$ in the denominator, rather than T. Nowadays, it is called Fisher's g test and is implemented in the `GeneCycle` package in `R` (Ahdesmaki et al. (2012)), for example.

Koen (1990) looks in the same problem as Fisher (1929), but in the context of astrophysics. However he shows that when we do not have observations at fixed interval or when we do not use the Fourier frequencies in calculating the periodogram, then the subtle difference between the exponential and F-distribution for the ratio of a given periodogram ordinate to the cumulative periodogram matters and this should be taken into account.

2.3.2 Bartlett's test for white noise

Bartlett (1950) further developed white noise testing by analysing general linear processes with continuous spectra. He also puts special emphasis of the work of Yule (e.g. Yule-Walker equations) with respect to the autoregressive process of order two. Bartlett's main motivation is the fact that while Fisher's test can show that there is a hidden periodicity, if there are more than one, then it is hard to say which one is the most important. Therefore, more elaborate methods are needed to perform the so-called "periodogram analysis" Bartlett (1950). Furthermore, his paper emphasizes the benefits of the general linear model and discusses general results for stationary time series such as Wiener-Khintchin theorem(page 2):

"*The very generality of these results becomes embarrassing when it comes to the analysis of actual series, and further assumptions about the character of the series are usually necessary in practice before much progress can be made — hence the importance of one general type of series distinct from the classical harmonic series and having a continuous spectrum.*"

Furthermore, Bartlett (1950) is remarkable since it employs many devices related to the spectral estimation from the periodogram, which are nowadays a standard in statistical time series software. For instance, Bartlett explains why it is useful to smooth the periodogram with the Daniell kernel when plotting it against frequency for spectrum visualization.

The argument (for smoothing) from Bartlett (1950) is the following. Let us have a univariate time series X_t, $t \in \mathbb{N}^+$ which contain a harmonic component with a fixed frequency λ, then its spectrum f_s would have a component $c \cos \lambda s$. Thus at frequency

$\omega_p = \lambda$ the periodogram's expectation $E(I_{\omega_p})$ will tend to infinity. For $\omega_p = \lambda$, this gives an $\mathcal{O}(T)$ effect to the calculated periodogram, where T is the length of the series X_t. On the other hand, for every $\omega_p \neq \lambda$, the effect is $\mathcal{O}(T^{-1})$ and the "periodogram analysis" can distinguish frequencies that have a difference in magnitude greater than this. Moreover, because of the derived exponential distribution of I_{ω_p} from Fisher (1929) i.e. $\mathbb{P}(I_{\omega_p} \geq z) = \exp\{-z/E(I_{\omega_p})\}$ and its memorylessness property, for an iid process, the periodogram alternates around the mean. Therefore, for spectral estimation Bartlett (1950) recommends that the series is smoothed, so that such fluctuations and the induced correlations among the periodogram ordinates are mitigated. It is well known today, that, asymptotically, the periodogram is not a consistent estimator of the spectrum i.e. its variance does not decrease when the number of observations T goes to infinity. As an aside, if we use the `spectrum` function in `R`, then smoothing and plotting the logarithm of the periodogram are activated automatically.

Moreover, the arguments developed in Bartlett (1950) help when one decides how to approach white noise testing in the spectral domain. Since we would not be looking for any particular frequency, or a combination of, then we are not restricted by the $\mathcal{O}(T^{-1})$ contributions. On the contrary, we might use methods for expressing deviation between statistical distributions in order to discern whether or not a time series exhibits (Gaussian) white noise characteristics. In chapter 4 we will be using distributional results for the wavelet coefficients of a periodogram in order to test for white noise.

Such approaches are developed in Bartlett (1954), Bartlett (1955) and Grenander and Rosenblatt (1957). The general idea is to evaluate the departure of the empirical periodogram from the one expected under the white noise hypothesis, accounting for the distributional differences. To accomplish this task, a Kolmogorov-Smirnov test is used, from Kolmogorov (1933) and Smirnov (1948). This test is based on the Glivenko-Cantelli theorem and considers the maximum difference of the empirical and theoretical distribu-

tion functions of a given data and theoretical distribution. The Bartlett test from Bartlett (1954) and Bartlett (1955), pages 92-94 is based on applying the Kolmogorov-Smirnov test to the cumulative distribution function(cdf) of the raw periodogram of the data, comparing it with a uniform cdf. To our knowledge, the Bartlett test is implemented in the commercial SAS and STATA statistical packages, and our `R` package `hwwntest`. Following Newton (1996), next we present the main steps of the Bartlett test for white noise.

Let us have a time series $\{X_t\}, \quad t = 1 \dots T, \quad T \in \mathbb{N}$ and $q = T/2$. First we define the cumulative (or integrated) periodogram as:

$$\hat{F}_I(\omega_p) = \frac{\sum_{j=1}^{p} I(\omega_j)}{\sum_{j=1}^{q} I(\omega_j)}, \quad p = 1 \dots q, \tag{2.12}$$

where $F_I(0) = 0$ and $F_I(\omega_q) = 1$. The idea of the Bartlett test is that, under the null hypothesis of $\{X_t\}$ being white noise, if we plot the integrated periodogram against the frequency, then the points should be placed on a straight line. The (Kolmogorov-Smirnov) test statistic is defined as:

$$B = q^{-1/2} \max_{1 \le k \le q} \left| \hat{F}(\omega_k) - \frac{k}{q} \right|, \tag{2.13}$$

and its distribution function is:

$$G(b) = \lim_{n \to \infty} Pr(B \le b) = \sum_{j=-\infty}^{\infty} (-1)^j \exp\{-2b^2 j^2\}. \tag{2.14}$$

Then, this distribution function G can be used to calculate a p-value.

A drawback of the Bartlett's test is that it does not perform well with small samples. The reason is that, the Kolmogorov-Smirnov test is based on the Glivenko-Cantelli central limit theorem which deals with the convergence of the empirical distribution function to the continuous distribution function. This theorem relies on the strong law of large numbers and thus implicitly requires a fine partition where the empirical cdf is evaluated i.e. a large number of observations. Thus, if we do not have many datapoints, then the

empirical cdfs would have more steep steps and large differences, since empirical cdfs are step functions with discontinuities.

2.3.3 Durbin's periodogram-based test

Durbin (1969) extends Bartlett's cumulative periodogram white noise test, considering the small-sample scenario as well as adapting the procedure to use least-squares residuals. Furthermore, it explores the scenarios of both high and low frequency departures from Gaussian white noise. Additionally, a graphical approach is presented which shows confidence bands for the respective high and low frequency departures from whiteness. A generalized test, based on the mean value of the integrated periodogram is also presented. Durbin (1969) stated that the reason for development of those tests is because the autocorrelation based Durbin-Watson tests(Durbin and Watson (1950, 1951) are not applicable to least-squares residuals and also, in practice, a better tool for detecting departures from white noise is needed due to the drawbacks of Bartlett's test.

When approaching the high versus low frequency departures from Gaussian white noise, Durbin (1969) considers two cases of the Bartlett's test statistic from equation (2.13), without the absolute value. For the situation when the alternative hypothesis is a low frequency departure, Durbin (1969) defines:

$$c^{+} = \max_{1 \le k \le q} \left\{ \hat{F}(\omega_k) - \frac{k}{q} \right\}. \tag{2.15}$$

For a high frequency alternative, Durbin (1969) defines:

$$c^{-} = \max_{1 \le k \le q} \left\{ \frac{k}{q} - \hat{F}(\omega_k) \right\}. \tag{2.16}$$

Eventually, Bartlett's test statistic is defined for the two-sided alternative hypothesis:

$$c = max(c^{+}, c^{-}) \tag{2.17}$$

The test procedure is thus very similar to the one from Bartlett (1954). However, the Durbin (1969) provides theoretical arguments as to what modifications in the confidence bands must be made when the residuals are from regressions on lagged dependent variables i.e. a simple autoregressive model. Durbin also points out that there are regions of the test statistics for which the results of the tests are inconclusive, similar to the ones from Durbin and Watson (1950) and Durbin and Watson (1951).

Furthermore, an additional procedure — for the average integrated periodogram value:

$$\overline{\widehat{F}_I}(\omega_p) = \frac{1}{q-1}\sum_{j=1}^{q-1}\hat{F}_I(\omega_p), \tag{2.18}$$

as a test statistic — is considered and also explained that its distribution would be simpler and the inconclusive region, such as the one for Durbin-Watson's d statistic from Durbin and Watson (1950), shown in equation 2.19, is narrower. The reason is that, because of the central limit theorem, the distribution of the average cumulative periodogram would converge quickly to the Normal distribution.

2.3.4 Other work related to Fisher's and Bartlett's tests

An interesting paper is Reschenhofer (1989) who looks separately at the sine and cosine components of the discrete Fourier transform(DFT). Reschenhofer (1989) uses the work of Fisher (1929) with respect to the F-distribution of the ratios of periodogram ordinates. The paper exploits the fact that Bartlett's white noise test often fails if there is more than one peak in the periodogram. Furthermore, it is shown through simulation that the autoregressive testing approach of Box and Pierce (1970) is superior to the approach in Reschenhofer (1989) when the lag tested is less than 5. However, for large lags such as 20, the Box-Pierce approach is inferior to the periodogram test developed in Reschenhofer (1989). It is also noted that the rank portmanteau test of Hallin et al. (1987), that is initially developed for stationary continuous alternative hypotheses, is also applicable for

compound periodicity alternatives where the periodogram has more than one peak.

2.3.5 Directions for periodogram-based tests and this thesis

From the works of Schuster (1898); Fisher (1929); Bartlett (1950, 1954); Durbin (1969), it is clear that the periodogram gives many opporunites for the development of white noise tests, both for raw datasets and residuals from regression (or other) models.

In this thesis, we use wavelet decompositions of the periodogram in chapter 4 in order to develop white noise tests that encompass different classes of alternative hypotheses and can be applied for either: high and low frequency or general alternatives with a continuous spectrum such as ARIMA models.

2.4 The autocorrelation approach to white noise testing

An approach for white noise testing based on the sample autocorrelation function is complementary to spectral analysis. From this perspective, the problem of white noise testing often arises when considering the residuals from linear regression. The reason is that, because of Gauss-Markov theorem, the statistics for overall model fit and coefficient significance such as the F and T would only be valid if the residuals of the model are iid. Thus, especially in econometrics, it became customary first to check whether or not the residuals from a model possess first-order autocorrelation, with the Durbin-Watson test (Kotz and Johnson (1992), page 234). It is probably interesting to notice that in the paper Durbin and Watson (1950) today's standard matrix notation for linear models was coined in i.e. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (Kotz and Johnson (1992), page 234).

2.4.1 The Durbin-Watson test for lag one serial correlation

One of the first tests available to search for residual autocorrelation was the Durbin-Watson test. The null hypothesis H_0 is that the residuals from a regression model are iid $N(0, \sigma^2)$. The procedure was developed in Durbin and Watson (1950) and statistical tables with critical values for the test statistic were produced in Durbin and Watson (1951). It concerns only lag one autocorrelation. The Durbin-Watson test statistic is:

$$d = \frac{\sum_{t=2}^{T}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{T}\epsilon_t^2}, \tag{2.19}$$

where ϵ_t are the residuals from a model such as $y_i = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1 \dots T, \quad T \in \mathbb{N}$. However, in practice, we do not know the real residuals, but only their estimates which will be denoted $\hat{\epsilon}_t$. So, considering lag one autocorrelation in the residuals of a model gives a regression equation from which, the autocorrelation ρ can be estimated as $\hat{\rho}$.

$$\hat{\epsilon}_t = \rho\hat{\epsilon}_{t-1} + v_t, \tag{2.20}$$

where v_t is an IID error term. Considering equations (2.19) and (2.20), and after some algebra, one can get the relationship between the two:

$$d = 2(1 - \hat{\rho}). \tag{2.21}$$

Thus, H_0 means that $\hat{\rho} = 0$ which translates to $d = 2$. However, the ease of use of this approximation depends on complicated arguments since there is a subtle difference between the observed and the real residuals. For a detailed review of the arguments, the reader is referred to either the articles Durbin and Watson (1950), Durbin and Watson (1951) or Kotz and Johnson (1992), page 229 for an extensive discussion. It would suffice to say that the performance of the Durbin-Watson test depends also on the number of parameters in the regression equation from which the residuals are derived. Furthermore, if we look closely at equation (2.21), we can notice that the test statistic d varies between

0 (perfect positive autocorrelation of residuals) and 4(perfect negative autocorrelation). This test statistic also has two different critical values — lower d_l and upper d_u — depending on whether the alternative hypothesis H_a is $\rho > 0$ or $\rho < 0$. So, there is a range of the test statistic for which the test is inconclusive. For example, given a fixed number of observations n, if the number of variables in the regression k increases, then the gap between d_l and d_u also increases, so the test loses in terms of empirical statistical power.

The Durbin-Watson test is ubiquitous in statistical packages — available in: `R`, package `lmtest`, function `dwtest`; Matlab, Mathematica, SAS, Eviews, Stata, Minitab, SPSS. For a review from applied point of view with the different steps in econometric analysis, consult Hatekar (2010), section 6.5.

2.4.2 The Extended Durbin-Watson h-test

A drawback of the Durbin-Watson(DW) test is that it is not applicable when the regressors are lagged values of the dependent variable. As such, it could not be applied for residual testing of autoregressive models, which is a large area for application of model validation technique such as the DW test. Especially, autoregressive model specification for the errors i.e. is the serial correlation in lag one or lag two. The problem was addressed by Durbin (1970) where a general likelihood ratio Neyman-Pearson type test is developed We will outline the test statistic, but not go into details of the method, since it is far from the methodology developed in this thesis. The test statistic for the extended DW test is:

$$h = (1 - d/2) \left[T / \{1 - T\widehat{\mathrm{Var}}(\widehat{\beta}_1\} \right]^{1/2}, \tag{2.22}$$

where T is the number of observations, d is the DW statistic from (2.19) and $\widehat{\mathrm{Var}}(\widehat{\beta}_1)$ is the estimated variance of the autoregressive parameter for lag one. Additional condition regarding the validity of the test is that $T \cdot \widehat{\mathrm{Var}}(\widehat{\beta}_1) < 1$ and the sample is moderately large (Hatekar (2010), section 6.5).

It is interesting to note that Durbin (1970) includes the Ljung-Box portmanteau test as a special case of the general methodology developed there.

2.4.3 Box-Pierce-Ljung white noise test

The Box-Pierce-Ljung test is one of the most popular white noise tests — consisting, formally, of two tests Box-Pierce and Ljung-Box. As with the Durbin-Watson test, it is available in most commercial and open source statistical packages. Unlike the DW test, it considers departures from white noise up to a pre-chosen lag and is thus a cumulative portmanteau test.

The Box-Pierce test is developed in Box and Pierce (1970) which derives the mathematical approximation of the distribution of residuals in autoregressive integrated moving average (ARIMA) models. The residual distribution is shown to be χ^2 with degrees of freedom equal to the lag, up to which the sample autocorrelations are tested. Let $X_t, \quad t = 1, \ldots, T$ be a zero mean second order stationary time series. Then $\hat{\rho}_i = \mathrm{Cov}(X_t, X_{t+i})/\mathrm{Var}(X_t)$ is the sample autocorrelation coefficient at lag i. The test Box-Pierce test statistic is:

$$T \sum_{i=1}^{h} {\hat{\rho}_i}^2, \tag{2.23}$$

where T is the number of observations and the index h denotes the maximum lag considered.

Later, it was realized that the approximation for the distribution of (2.23) could be improved and Ljung and Box (1978) was developed. The Ljung-Box test statistic is:

$$Q = T\,(T+2) \sum_{i=1}^{h} \frac{\hat{\rho}_i^2}{T-i}, \tag{2.24}$$

where T is the sample size, $\hat{\rho}_i$ is the sample autocorrelation at lag i, h is the number

of lags being tested and Q is approximately $\chi^2(h)$ distributed. However, because of the differences between the residuals and true white noise autocorrelations, when considering the residuals from an ARIMA$(p, 0, q)$ model, the degrees of freedom should be corrected: $Q \sim \chi^2(h - p - q)$.

2.4.4 Ljung-Box evaluation and lag selection

Due to its popularity and ease of computation, the Ljung-Box test is often the first choice of a practitioner and is ubiquitous in statistical software. However, there exists the crucial question of selecting the lag. It can be seen, from its test statistic in (2.24), that the higher the lag, the larger the test statistic. Moreover, this means that choosing a high lag could drive the power of the test to deteriorate if the true statistically significant autocorrelation coefficient is contained in smaller lags. On the other hand, choosing a small lag could miss significant higher-lag autocorrelation coefficients. A survey of its statistical power and rules of thumb for lag selection are described by Hyndman (2014). Hyndman (2014) employs a Kolmogorov-Smirnov test for the uniformity of the p-values calculated from simulated data. The main conclusion is that when we increase the lag, then the statistical size becomes greater than the nominal size of the test and this is checked for nominal levels of both 5% and 1%.

2.5 Contemporary work on white noise testing

With the increasing data abundance and the development of the financial markets in the 1980s, many more time series models were developed. Examples are ARCH,Engle (1982), and GARCH, Bollerslev (1986), which model the variance (volatility) of a time series as an ARIMA process itself. These developments created a new wave in econometric applications of white noise tests over the past 25 years. A detailed introduction is available in Guay et al. (2013) and Lobato et al. (2001). For our exposition, it suffices to explain what are the main directions of some recently developed tests.

2.5.1 Small magnitude autocorrelation

The origin of this econometric problem arises in financial time series. In plain terms, the efficient market hypothesis means that financial markets are effective and thus there should be no autocorrelation in the financial returns time series. However, this is not always true, so institutions can make money on the market, Samuelson (1965). The zero autocorrelations phenomenon is also known as one of the stylized facts for financial time series and one reason for the development of the ARCH/GARCH models. It is an important example because near-zero sample autocorrelation is on the edge of the Box-type tests. To illustrate that, let us think about the standard confidence intervals for the autocorrelations, under the null hypothesis of independent and identically distributed time series: $-2/T^{-1/2}$ to $+2/T^{-1/2}$ where T is the number of observations. If we have a moderately long series, say $T = 200$ to 300 e.g. $T = 256$, then our 95% confidence interval would be from $-2/16$ to $2/16$, i.e. an autocorrelation of 0.125 would not be deemed significant. A similar test scenario is used in Guay et al. (2013), namely, to be able to detect an $o(T^{-1/2})$ autocorrelation, provided there is enough of them. Their way to achieve that is to use a heavily-modified Box-Pierce test, more precisely, weighting its autocorrelations via special kernels and/or tuning algorithm for choosing the right lag. For instance, the Box-Pierce might be considered as a vanilla case with uniform kernel and pre-specified lag. These procedures are by no means simple and would be a huge burden for a practitioner.

A test based on periodogram/weighted spectral density. which covers the ARFIMA alternative and local AR(1) and MA(1) alternatives, is developed in Hong (1996). It is based on kernel-weighted distance (metric) between the null-hypothesized spectral density and the estimated one from the data. The author shows three variants: one with quadratic form distance, one based on Hellinger metric and one based on information criterion.

2.5.2 The tests from Guay et al. (2013) and Hong (1996)

The Guay et al. (2013) and Hong tests merge the spectral approach, explained earlier in section 2.3, with the Box-Pierce-Ljung approach. The assumptions of Hong (1996) are that we have a second-order stationary process $\{X_t\}$, with an autocorrelation function $\rho_j, \quad j \in \mathbb{Z}$ However, instead of starting with the periodogram, Hong (1996) starts with a normalized spectral density function:

$$f(\omega) = (2\pi)^{-1} \sum_{-\infty}^{\infty} \rho_j \cos j\omega, \quad \omega \in [-\pi, \pi]. \tag{2.25}$$

The null hypothesis H_0 is that $\rho_j = 0$ for all $j \in \mathbb{Z}, \quad j \neq 0$ and the alternative H_a is that $\rho_j \neq 0$ for some $j \neq 0$. Let also $f_0(\omega) = 1/2\pi$ be the spectrum of white noise. Hong (1996) considers different divergence measures to express the deviation of f from f_0 — based on quadratic norm, Hellinger metric and Kullback-Leibler divergence. We will explain his test when using the quadratic norm, since the situation is similar for the others. In the quadratic norm case the test statistic is:

$$Q(f, f_0) = \left[2\pi \int_{-\pi}^{\pi} \{f(\omega) - f_0(\omega)\}^2 d\omega \right]^{1/2}. \tag{2.26}$$

Then the test statistic can be computed as the value of $Q(\hat{f}_t, f_0)$ where $\hat{f}_t$ is a kernel estimator for the spectral density from the data. Then it is explained, that the Box-Pierce test can be seen as a special case of $Q(f, f_0)$ where f is a truncated periodogram. Hong (1996) goes further by considering kernels and theoretical arguments for the selection of a *lag* h, which would determine the magnitude of the weights for the kernel density estimation of the spectrum. Additionally, scaling quantities are defined to modify the statistic Q, together with assumptions, which lead to a form that has asymptotic Gaussian distribution, from which the actual test is constructed.

Guay et al. (2013) uses the test statistic Q as a starting point, however then develops complicated algorithms for lag selection and weighting in order to fullfil the case for the alternative hypothesis of the type $o(T^{-1/2})$. We will use scenarios from both Guay et al. (2013) and Hong (1996) to compare to results of our wavelet-based white noise tests in chapter 4. However, we could not find a software implementation of the Guay et al. and Hong tests.

2.5.3 Lobato-Velasco periodogram-based test

Lobato and Velasco (2004) introduced a spectral white noise test based on the standardized cumulative periodogram, in fact the statistic that is maximized from Bartlett's test. As defined before, let us have a time series $X_t, \quad t = 1 \dots T, \quad T \in \mathbb{N}$, $q = T/2$ and define.

$$Z(\omega_p) = q^{-1/2} \left\{ \hat{F}_i(\omega_p) - \frac{k}{q} \right\}, \tag{2.27}$$

where $\hat{F}_i$ is the cumulative periodogram from section 2.3.2, equation (2.12)

A key component in this test is also the Cramer von Mises statistic given by:

$$CvM = q^{-1} \sum_{p=1}^{q} Z(\omega_p)^2 \tag{2.28}$$

Let us have a second-order stationary time series $\{X_t\}$ with mean $\mu = 0$ and variance $\sigma^2 < \infty$. Next, Lobato and Velasco (2004) reformulate the null hypothesis in terms of the periodogram — since under white noise the normalized spectrum would be 1, then $I(\omega)/\hat{\sigma^2} - 1 = 0$ and then the Cramer von Mises functional is applied to it which gives:

$$M_n = \frac{1}{q} \sum_{p=1}^{q} \left\{ \frac{I(\omega_p)}{\hat{\sigma^2}} - 1 \right\}^2 \tag{2.29}$$

Further in the paper, it is shown that $q^{-1/2}(M_n - 1)$ is asymptotically distributed as

$N(0, 4)$ and it is also explained that the test is suitable for analysing regression residuals. The `R` package `normwhn` contains the function `whitenoise.test` which computes the Lobato-Velasco test. We will use this test for comparisons in chapter 4 of the thesis.

2.5.4 Overview of white noise tests assumptions

Another important aspect underlying the construction of a statistical test are the assumptions which are made. In the tests of Bartlett, Durbin's both periodogram based and autocorrelation of lag one tests, and Box-Pierce-Ljung tests, the main assumptions for H_0 are second order stationarity and that the process X_t is Gaussian. However, with the knowledge of Proposition 1, the Gaussian assumption in Bartlett's test may be relaxed.

In Reschenhofer's and our wavelet tests developed in chapter 4, only second order stationarity is assumed.

In contrast, due to the more complicated setups and scenarios, the contemporary tests of Hong, Guay and Lobato require more stringent assumptions (additional to second order stationarity) that are hard to verify for a real time series data:

1. The test of Hong assumes only IID data, but requires a finite moment condition of order four.

2. The test of Guay et al. (2013) assumes a finite moment condition of order 12 and the order of the average lag may reach $T^{1/2}$.

3. The test of Lobato and Velasco (2004) assumes a finite moment condition of order eight.

2.5.5 White noise tests summary and thesis directions

The literature appears to contain two main approaches to the white noise hypothesis testing: the spectral and sample autocorrelation approach. The key tests are Bartlett's and

Box-Pierce-Ljung. There is a plethora of other tests which represent different modifications of those two main tests or tuning parameters.

We will be using wavelet transformations of the periodogram in order to construct our white noise tests in chapter 4. The variety of wavelets gives us the opportunity to construct tests which suit different alternative hypotheses, without changing the test statistics as in Durbin (1969), but by using different wavelet specifications.

2.6 Wavelets and wavelet transforms

Wavelets are mathematical functions that are "little waves" and they are often used to represent other functions. In mathematical sense they are quite new, since the theory has been developed predominantly in the last 30 years. Some of the most comprehensive mathematical reference to wavelets are Daubechies (1992), Mallat (1998), Vidakovic (1999), Percival and Walden (2000) or Nason (2008), the last 2 dealing with wavelets in time series analysis and statistics. Similar to Fourier series, wavelets can act as a basis for $L^2(\mathbb{R})$ space of square-integrable functions on the real line. Unlike Fourier series' basis functions, wavelets can be approximated by functions with compact support. Furthermore, due to their construction, wavelets provide *sparse* representations for smooth functions. For instance, if a smooth function has some discontinuities, then only a few wavelet coefficients of its representation would be influenced. However, if we calculate the Fourier coefficients of a function with one discontinuity, then all coefficients would be influenced since the support of the sine and cosine is the whole real line (e.g. the Gibbs phenomenon).

Furthermore, there exist many different wavelet bases that carry certain properties with respect to simulaneous localization in time and frequency. Moreover, there are orhogonal wavelet bases, but also non-orthogonal ones, thus wavelets can be used for

different purposes, especially when analysing nonstationary signals and creating models for them (Nason et al. (2000)). Similarly to Fourier analysis, there can be a wavelet spectrum which, however, would contain information for the variance over the scale rather than frequency of the data e.g. we will have a scalogram rather than a periodogram (Nason and von Sachs (1999), Percival (1995)). In order to illuminate on those concepts, first we will define and show the oldest and simplest type of wavelets — Haar — in the next section.

2.6.1 Haar wavelets in one dimension

The coverage of the $L^2(\mathbb{R})$ space of functions from wavelets is achieved through the translation and scaling(dilation) of a single function called the mother wavelet. Let us now describe the one-dimensional Haar wavelet system.

Define

$$\psi(x) = \begin{cases} 1, & \text{if } x \in [0, 1/2) \\ -1, & \text{if } x \in [1/2, 1) \\ 0, & \text{otherwise} \end{cases} \tag{2.30}$$

to be the mother wavelet — a function $\psi : \mathbb{R} \mapsto \mathbb{R}$. Next, the wavelets are defined through the mother Haar wavelet:

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z} \tag{2.31}$$

The collection $\{\psi_{j,k}(x)\}$ forms an orthonormal basis for $L^2(\mathbb{R})$. Therefore, given a function $f \in L^2(\mathbb{R})$ we can write:

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(x), \tag{2.32}$$

where

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)dx. \tag{2.33}$$

Equations (2.32) and (2.33) are called the wavelet transform and $d_{j,k}$ are the wavelet coefficients of a function $f(x)$. Equation (2.33) is simply the (scalar) dot product in $L^2(\mathbb{R})$ i.e. $d_{j,k} =< f(x), \psi_{j,k}(x) >$. Based on the allowed ranges for j and k, wavelets can be discrete or continuous. Furthermore, j is called the scale and k the location.

Next, for illustration, let us see how the discrete Haar wavelets are formed from relations (2.30) and (2.31). Let $j \in \mathbb{N}$ and define:

$$\psi_{j,k} = \begin{cases} 2^{-j/2}, & \text{if } k = 0, \ldots, 2^{j-1} - 1 \\ -2^{-j/2}, & \text{if } k = 2^{j-1}, \ldots, 2^j - 1 \\ 0, & \text{otherwise} \end{cases}$$

So, we also notice that the number of locations depends on the number of scales:

$$\psi_{1,k} = 2^{-1/2}[1, -1]$$
$$\psi_{2,k} = 2^{-1}[1, 1, -1, -1]$$
$$\psi_{3,k} = 2^{-3/2}[1, 1, 1, 1, -1, -1, -1, -1]$$

2.6.2 How wavelets are constructed

How Haar wavelets are generated

As explained in section 2.6.1, the mother wavelet determines the Haar wavelet orthogonal system. However, this is not the whole story and there is also a father wavelet function which, transformed in a particular way, defines the mother wavelet. In this section we will explain this process by following Walter (1994) chapter 1.

The father wavelet of the Haar orthogonal system is the indicator function of the $[0, 1)$ interval of the real line, defined by:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in [0, 1) \\ 0, & \text{otherwise} \end{cases} . \tag{2.34}$$

Then it immediately follows that $\phi(x)$ and $\phi(x - k)$, $k \neq 0$, $k \in \mathbb{Z}$ are orthogonal because their product is zero. However, $\{\phi(x - k)\}$ is not an orthogonal system for $L^2(\mathbb{R})$ since its closed linear system V_0 contains piecewise constant functions with possible jumps only at the integers, but the integers are not dense in $\mathbb{R}$.

Therefore, in order to include more functions, a dilated version of $\phi(x)$ can be defined as $\phi(2^j x)$, $j \in \mathbb{Z}$. Using a change of variable, it can be shown that $\{2^{j/2}\phi(2^j x - k)\}$ is a complete system. Let its closed linear span be denoted by V_j. Those linear subspaces play a major role in constructing other wavelet bases as well. $\bigcup_j V_j$ is dense in $L^2(\mathbb{R})$, because any function in $L^2(\mathbb{R})$ can be approximated by a piecewise constant function f_j with jumps at binary rationals. Consequently, the system $\{\phi_j\}$, defined by:

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k),$$

is complete in $L^2(\mathbb{R})$, but not orthogonal because $\phi(x)$ and $\phi(2x)$ are not orthogonal. However, this issue might be circumvented by defining

$$\psi(x) = \phi(2x) - \phi(2x - 1). \tag{2.35}$$

Now, $\{\psi(x-k)\}$ is an orthonormal system and $\psi(2x-k)$ and $\psi(x-k)$ are orthogonal

for all x and k. Therefore, it follows that the system $\{\psi_{jk}\}_{j,k\in\mathbb{Z}}$ defined by:

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

is complete and orthonormal in $L^2(\mathbb{R})$. We recognize that this is equation (2.31). Also, considering equation (2.32) and the wavelet coefficients in equation (2.33) (calculated as a dot product in $L^2(\mathbb{R})$ i.e. $< f, \psi_{jk} >$) , we can write for any function $f \in L^2(\mathbb{R})$:

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} < f, \psi_{jk} > \psi_{j,k}(x). \tag{2.36}$$

Then the wavelet approximation of a function f, for a scale j, is given by:

$$f_j(x) = \sum_{i=-\infty}^{j-1} \sum_{k=-\infty}^{\infty} < f, \psi_{ik} > \psi_{i,k}(x), \tag{2.37}$$

which converges to a piecewise constant function with jumps at $2^{-j}x$, $x \in \mathbb{Z}$ for Haar wavelets.

2.6.3 The Shannon wavelet and multiresolution analysis

We continue to follow Walter (1994) chapter 1 for our exposition.

We started section 2.6.2 with the indicator function of the unit interval which played the role of the father wavelet. The father wavelet is also called scaling function in the wavelet literature. The Shannon wavelet system is complementary to the Haar system in the sense that it starts with the indicator function in the frequency domain, whereas in Haar, we started in the time domain. Therefore let us call it the Fourier transform of the scaling function, defined to be:

$$\hat{\phi}(\omega) = \begin{cases} 1, & \text{if } -\pi \leq \omega \leq \pi, \\ 0, & \text{otherwise.} \end{cases} \tag{2.38}$$

By fundamentals, its inverse Fourier transform is:

$$\phi(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} \hat{\phi}(\omega) \exp(i\omega x) d\omega = (2\pi)^{-1} \int_{-\pi}^{\pi} \exp(i\omega x) d\omega = \sin \pi x / \pi x. \quad (2.39)$$

$\phi(x)$ is called the *sinc* function, its value at $x = 0$ is defined as $sinc(0) = 1$, following from application of L'Hopital's rule for $x \to 0$. Here $\phi(x)$ and $\phi(x-k)$, $k \in \mathbb{Z}$ are orthogonal based on Fourier transform properties and Parseval's inequality. It can be shown that:

$$\int_{-\infty}^{\infty} \phi(x)\phi(x-k)dx = \sin(\pi k)/\pi k = 0, \quad k \neq 0.$$

If we have a function $f(x) \in L^2(\mathbb{R})$ which has a Fourier transform $\hat{f}(\omega)$ that vanishes for $|\omega| > |\pi|$, it has a Fourier series given by

$$\hat{f}(\omega) = \sum_k c_k \exp(i\omega k), \quad |\omega| \leq \pi, \quad (2.40)$$

where c_n are the Fourier coefficients. Then, by the Fourier inversion theorem applied to both sides of (2.40), it can be shown that:

$$f(x) = \sum_k f(-k) \sin \pi(x+k)/k(x+k), \quad \forall x. \quad (2.41)$$

Let V_0 denotes the space of all such functions $f(x)$. It is a closed linear space. Formula (2.41) is called the Shannon sampling theorem. It is used to reconstruct a band-limited function from V_0 from its realizations over the integers. If we make a change of variable in $f(x)$ from equation (2.41) — $x = 2t$ and get $g(t) = 2t$, then we have a new closed linear space V_1 that consists of functions with Fourier transforms vanishing outside $[-2\pi, 2\pi]$. This process can be repeated infinitely many times in both shrinking and expanding change-of-variables. As a result we get a sequences of subspaces $\{V_j\}_{j=-\infty}^{\infty}$

that are ordered in the following relation:

$$\cdots \subseteq V_{-j} \subseteq \cdots \subseteq V_{-1} \subseteq V_0 \subseteq V_1 \subseteq \cdots V_j \subseteq \cdots \tag{2.42}$$

When $j \to -\infty$ in (2.42), the support of the Fourier transform of ϕ collapses to zero. When $j \to \infty$, it covers the whole real line. So, $\{V_j\}_{j=-\infty}^{\infty}$ has the following key properties:

1. $\bigcap_j V_j = \{0\}$

2. Every $f \in L^2(\mathbb{R})$ can be approximated by a function in V_j for a sufficiently large j

The sequence $\{V_j\}$ is called a "multiresolution analysis" associated with $\phi(x)$.

Now, similarly to section 2.6.2, relation 2.35, we can construct the mother wavelet of the Shannon system. It is a function $\psi(x) \in V_1$ that is orthogonal to $\psi(x-k)$ and given by:

$$\psi(x) = 2\phi(x) - \phi(x), \tag{2.43}$$

whose Fourier transform is:

$$\hat{\psi}(\omega) = \hat{\psi}(\omega/2) - \hat{\psi}(\omega) \quad \omega \in [-2\pi, -\pi] \cup [\pi, 2\pi]. \tag{2.44}$$

Because the supports of $\hat{\phi}$ and $\hat{\psi}$ are disjoint, they are orthogonal. So, the inverse Fourier transform ψ is a mother wavelet for the Shannon system and generates an orthonormal sequence as for the Haar system in section 2.6.2. An important fact is that the requirement for a function f to be represented by Shannon wavelets is only to have a Fourier transform $\hat{f} \in L^1(\mathbb{R})$, and not compactly supported.

2.6.4 Multiresolution Analysis

In the previous section we showed how multiresolution analysis (MRA) is developed naturally for definition of orthonormal wavelets. This concept has been developed by Mallat (1989) and Meyer (1993). As also outlined in the construction of the Shannon wavelet in section 2.6.3, MRA helps to decompose a function or signal into non-overlapping frequency bands. Therefore, it can be used for developing different filters. Moreover, MRA provides the mechanism for creating the *discrete wavelet transform* and constructing an orthonormal basis for $L^2(\mathbb{R})$.

Definition 4. *Multiresolution analysis(Mallat, 1989)*
A MRA of $L^2(\mathbb{R})$ is a sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$, such that for every $j, k \in \mathbb{Z}$ (k is called the location and j is called the scale):

1. $V_j \subset V_{j+1}$
2. $\bigcap_j V_j = \{0\}$
3. $\bigcup_j V_j = L^2(\mathbb{R})$
4. $f(x) \in V_0 \Leftrightarrow f(2^j x) \in V_j$
5. $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$
6. *There exists a scaling function (father wavelet) $\phi \in V_0$ with $\int_{-\infty}^{\infty} \phi(x)dx = 1$ such that $\{\phi_{0,k} := \phi(t-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis in V_0*

We have already seen in the previous section that any function $f \in L^2(\mathbb{R})$ can be approximated by its projections onto the spaces V_j — this is ensured by properties 1, 2 and 3. Properties 4 and 5 allow that:

$$\phi_{j,k}(z) := 2^{j/2}\phi(2^j x - k) \quad k \in \mathbb{Z} \tag{2.45}$$

is an orthonormal basis for V_j, $j \in \mathbb{Z}$. The functions $\phi_{j,k}$ are called translations and dilations of the scaling function ϕ. Finally, property 6 ensures that any function $f \in V_0$ can be represented as a linear combination:

$$f = \sum_k c_k \phi_{0,k},$$

where c_k is the dot product in $L^2(\mathbb{R})$ given by:

$$c_k =< f, \phi_{0,k} >= \int_{-\infty}^{\infty} f(x)\phi(x-k)dx.$$

Let P_j denote the projection on V_j:

$$(P_j f)(x) = \sum_k c_{j,k}\phi_{j,k},$$

where $c_{j,k}$ are the scaling coefficients calculated by

$$c_{j,k} =< f, \phi_{j,k} >= \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)dx. \tag{2.46}$$

Equation(2.46) shows that $\forall f \in L^2(\mathbb{R})$ can be approximated by the elements of the subspaces V_j and properties 2 and 3 of MRA mean that the quality of approximation increases as $j \to \infty$.

Note on the filters h and g

Using the MRA from Definition (4), the scaling function or father wavelet ϕ itself can be represented from finer spaces V by wavelet coefficients similar to (2.46). Since $\phi \in V_0 \subset V_1$, we have $\phi(x) = \sum_{k\in\mathbb{Z}} h_k \phi_{1,k}(x)$ and $h_k =< \phi, \phi_{1,k} >$ because $\{\phi_{1,k}\}$ is an orthonormal basis for V_1. Then we have the relation:

$$\phi(x) = 2^{1/2} \sum_{k\in\mathbb{Z}} h_k \phi(2x-k) \tag{2.47}$$

Equation (2.47) is called the scaling or dilation equation and h_k is called a *low-pass filter* associated with ϕ. This relation can be generalized for any V_j and V_{j-1} from an MRA by:

$$\begin{aligned} < \phi_{j-1,k}, \phi_{j,k} > &= \int \phi_{j-1,k}(x)\phi_{j,k}(x)dx \\ &= \int 2^{1/2}\phi(t)\phi(2t+2k-n)dt \quad t = 2^{j-1}x - k \\ &= h_{n-2k} \end{aligned} \tag{2.48}$$

and

$$\begin{aligned} \phi_{j-1,k} &= \sum_{n\in\mathbb{Z}} < \phi_{j-1,k}, \phi_{j,n} > \phi_{j,n}(x) \\ &= \sum_{n\in\mathbb{Z}} h_{n-2k}\phi_{j,n}(x), \end{aligned} \tag{2.49}$$

which is called the scaling function refinement relation.

Next we will explain how mother wavelet functions are constructed from an MRA. Let W_j denote the orthogonal complement of V_j in V_{j+1}. Because V_j is a sequence of closed subspaces, we have $V_{j+1} = V_j \oplus W_j$ which holds for all $j \in \mathbb{Z}$. So, for a $j > J$ we get:

$$V_{j+1} = V_J \oplus \bigoplus_{k=0}^{j-J} W_{J-k}, \tag{2.50}$$

which means that a function in V_j can be represented by the sum of its approximation at a lower scale J and the 'detail' lost by going from scale J to j. Then, by using properties 2 and 3 from Definition 4, we have:

$$L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j. \tag{2.51}$$

Therefore, each function $f \in L^2(\mathbb{R})$ can be partitioned over W_j subspaces by its orthogonal projections. Moreover, those subspaces have the scaling property 4 from Definition

4 i.e.

$$f(x) \in W_j \Leftrightarrow f(2^j x) \in W_j, \quad \forall j \in \mathbb{Z}.$$

Furthermore, for a $\psi \in W_0$ we have the analogue of the scaling equation (2.47):

$$\psi(x) = 2^{1/2} \sum_{k \in \mathbb{Z}} g_k \phi(2x - k), \tag{2.52}$$

where the g_k is called the high-pass filter associated with the wavelet function ψ (see Daubechies (1992)). Moreover, the filters h_k and g_k are called *quadrature mirror filters* and are related by Daubechies (1992) page 326:

$$g_k = (-1)^k h_{1-k}, \quad k \in \mathbb{Z}. \tag{2.53}$$

Now, recalling from Definition 4 and equation (2.45), we can define similarly:

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k), \quad j, k \in \mathbb{Z} \tag{2.54}$$

then we have the following chained relations: the set $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ is a basis for $L^2(\mathbb{R})$; $\{\psi_{j,k}\}_{k \in \mathbb{Z}}$ i.e. for a fixed j is an orthonormal basis for W_j. Hence $\{\psi_{0,k}\}_{k \in \mathbb{Z}}$ is an orthonormal basis for W_0.

Similarly to equation (2.49), a relation for the mother wavelet function can be shown:

$$\psi_{j-1,k}(x) = \sum_{n \in \mathbb{Z}} g_{n-2k} \phi_{j,n}(x). \tag{2.55}$$

Using (2.55), (2.49) and (2.53), the filters can be written as:

$$h_{n-2k} = < \phi_{j,k}, \phi_{j+1,n} > \quad \text{and} \quad g_{n-2k} = < \psi_{j,k}, \phi_{j+1,n} > . \tag{2.56}$$

If the scaling function has compact support (e.g. the Haar wavelet), the actual coefficients

of the filters from equation (2.56) can be calculated by taking $k = j = 0$. Moreover, in the case of compactly supported scaling function the filters h and g have a finite number of non-zero coefficients and the mother wavelet ψ can be represented by a finite linear combination of functions with compact support.

Let $\delta_{a,b} = 1$ if $a = b$ and 0 for $a \neq b$ be the Kronecker delta function. Vidakovic (1999) shows that the filters have the following orthogonality properties:

$$\sum_{n\in\mathbb{Z}} h_n h_{n-2k} = \delta_{0,k} \quad , \quad \sum_{n\in\mathbb{Z}} g_n g_{n-2k} = \delta_{0,k} \quad , \quad \sum_{n\in\mathbb{Z}} h_n g_{n-2k} = 0 \tag{2.57}$$

2.6.5 The discrete wavelet tranform(DWT)

Function Representation given a Multiresolution Analysis

Section 2.6.1 shows that the wavelet transform in (2.32) and the $L^2(\mathbb{R})$ representation of a given function via wavelet coefficients in (2.33). Given a MRA for $L^2(\mathbb{R})$ we know that $\{\phi_{j,k}(x)\}_k$ and $\{\psi_{j,k}(x)\}_k$ form orthonormal bases for V_j and W_j respectively. Then for a fixed j_0, the collection $\{\phi_{j_0,k}(x)\}_{k\in\mathbb{Z}} \cup \{\psi_{j,k}(x)\}_{j>j_0,k\in\mathbb{Z}}$ forms an orthonormal basis for $L^2(\mathbb{R})$. Hence, $\forall f \in L^2(\mathbb{R})$ can be written as:

$$f(x) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(x) + \sum_{j>j_0}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x), \tag{2.58}$$

where $c_{j_0,k} =< f, \phi_{j_0,k} >$ and $d_{j,k} =< f, \psi_{j,k} >$, because of relations (2.51), (2.50) and property 2 of MRA from Definition (4). Furthermore, due to relation (2.51), $\forall f \in L^2(\mathbb{R})$ can be written as

$$f(x) = \sum_{j\in\mathbb{Z}} d_{j,k}\psi_{j,k}, \tag{2.59}$$

where the wavelet coeffficients $d_{j,k}$ provide the approximation of the function at scale j and location $2^{-j}k$.

Mallat's algorithm for DWT

A fast and efficient way to compute both $c_{j,k}$ and $d_{j,k}$ without calculating inner products at each scale and location is the algorithm is due to Mallat (1989) described as follows. From the filter relation (2.56) and $V_j \perp W_j$, it can be deduced that:

$$c_{j-1,k} = \sum_n h_{n-2k} c_{j,n} \tag{2.60}$$

and

$$d_{j-1,k} = \sum_n g_{n-2k} c_{j,n}, \tag{2.61}$$

and let the sets of coefficients be denoted as $\mathbf{c}_j = \{c_{j,k}\}_{k\in\mathbb{Z}}$ and $\mathbf{d}_j = \{d_{j,k}\}_{k\in\mathbb{Z}}$. Equations (2.60) and (2.61) imply that coefficient vectors for coarser scales $\mathbf{c}_{j-1}$ and $\mathbf{d}_{j-1}$ can be obtained from finer scale ones. That is a very important point practically, since in statistics we usually have data which might be treated as a discretized function. For instance, let us have a realization of a discrete stochastic process $X_t, t = 1, \dots T$,and $T = 2^J, J \in \mathbb{Z}^+$ might be an observation of a function let $X_t = f(x_t)$. We will use equation (2.58) with $J = j_0$ which implies that $c_{J,t} = f(x_t)$ is the finest possible representation. Next, using the filters h and g as in (2.60) and (2.61) recursively we can calculate all possible coefficients, given a fixed finest scale J. At each step, the number of coefficients is halved i.e. starting with $\#(c_J) = 2^J$, the first iteration yields $\mathbf{c}_{J-1}$ and $\mathbf{d}_{J-1}$ whose count is 2^{J-1} and so on to the last coefficient 2^0, i.e. $\mathbf{c}_0$. Conversely, the discrete wavelet transform can be inverted by the process of calculating finer scale coefficients from coarser scale ones recursively using:

$$c_{j,k} = \sum_{l=0}^{2^{j-1}} c_{j-1,l} h_{k-2l} + \sum_{l=0}^{2^{j-1}} d_{j-1,l} g_{k-2l} \tag{2.62}$$

Filter notation and decimation

The h and g filters from equation (2.56) can be thought as discrete convolution operators $\mathcal{H}$ and $\mathcal{G}$ on a sequence s_n:

$$(\mathcal{H}s) = \sum_n h_{n-k} s_n \tag{2.63}$$

and

$$(\mathcal{G}s) = \sum_n g_{n-k} s_n, \tag{2.64}$$

respectively. We also define the decimation operator $\mathcal{D}_0$ which picks the even elements out of a sequence s_n:

$$(\mathcal{D}_0 s)_k = s_{2k} \tag{2.65}$$

Using those operators, (2.60) and (2.60) can be written as:

$$c_{j-1,k} = \sum_n h_{n-2k} c_{j,n} = \{\mathcal{D}_0(\mathcal{H}c_{j,n})\}_k \tag{2.66}$$

and

$$d_{j-1,k} = \sum_n g_{n-2k} c_{j,n} = \{\mathcal{D}_0(\mathcal{G}d_{j,n})\}_k, \tag{2.67}$$

and this is how the discrete wavelet transform is implemented in practice. Here we described the *decimated* discrete wavelet transform which is orthogonal. However it is not translation-invariant, meaning that if we start at a different position in the sequence of datapoints we will get a different result. If we do not use the decimation operator $\mathcal{D}_0$ in (2.67) and (2.66), we will end up with the *non-decimated wavelet transform* which is not orthogonal, but is translation equivariant. For more information see Nason and Silverman (1994) or Coifman and Donoho (1995).

2.6.6 Vanishing moments

Definition 5. *A wavelet $\psi(x)$ has m vanishing moments if*

$$\int x^{\ell}\psi(x)\,dx = 0, \quad \ell = 0, \ldots, m-1, \quad m \in \mathbb{Z}^{+} \tag{2.68}$$

This property means that when a wavelet transform is applied to a polynomial of a degree less than m, then many of the $d_{j,k}$ coefficients will be exactly zero.

Daubechies extremal phase wavelets

The vanishing moments property in Definition 5 is a defining property for deriving Daubechies extremal phase wavelets, denoted D_n. The theory is developed in Daubechies (1992) chapters 6 and 7. As in section 2.6.2, wavelet construction starts with the scaling function ϕ from which the mother wavelet is then derived. However, in the case of Daubechies extremal phase wavelets, the scaling function does not have a closed-form expression, except for the Haar wavelets, which corresponds to the case of one vanishing moment. The support of the scaling functon and the mother wavelet depend on the number of vanishing moments m. If wavelet is to have $m > 1$ vanishing moments, then the (compact) support of the scaling function is to be $[0, 2m-1]$ and the mother wavelet's $[1-m, m]$ respectively. Thus if we have $m = 1$, we get the $[0, 1]$ support for the Haar father and mother wavelts. The actual father and mother wavelets are constructed by a cascade algorithm using the filters h and g, optimising the extremums with respect to the phase of the filters, taking into account the orthogonality conditions for the vanishing moments i.e. if we have m vanishing moments this would give 2^{m-1} equations to solve simultaneously. Then this cascade algorithm yields the h and g filter coefficients.

For wavelet computation and creating figures in this thesis we use the `wavethresh` package from **?**. Fig. 2.1 shows Daubechies Extremal Phase wavelets with 1, 5 and 10 vanishing moments. Looking at the graph, it is now clearer why wavelets represent little waves.

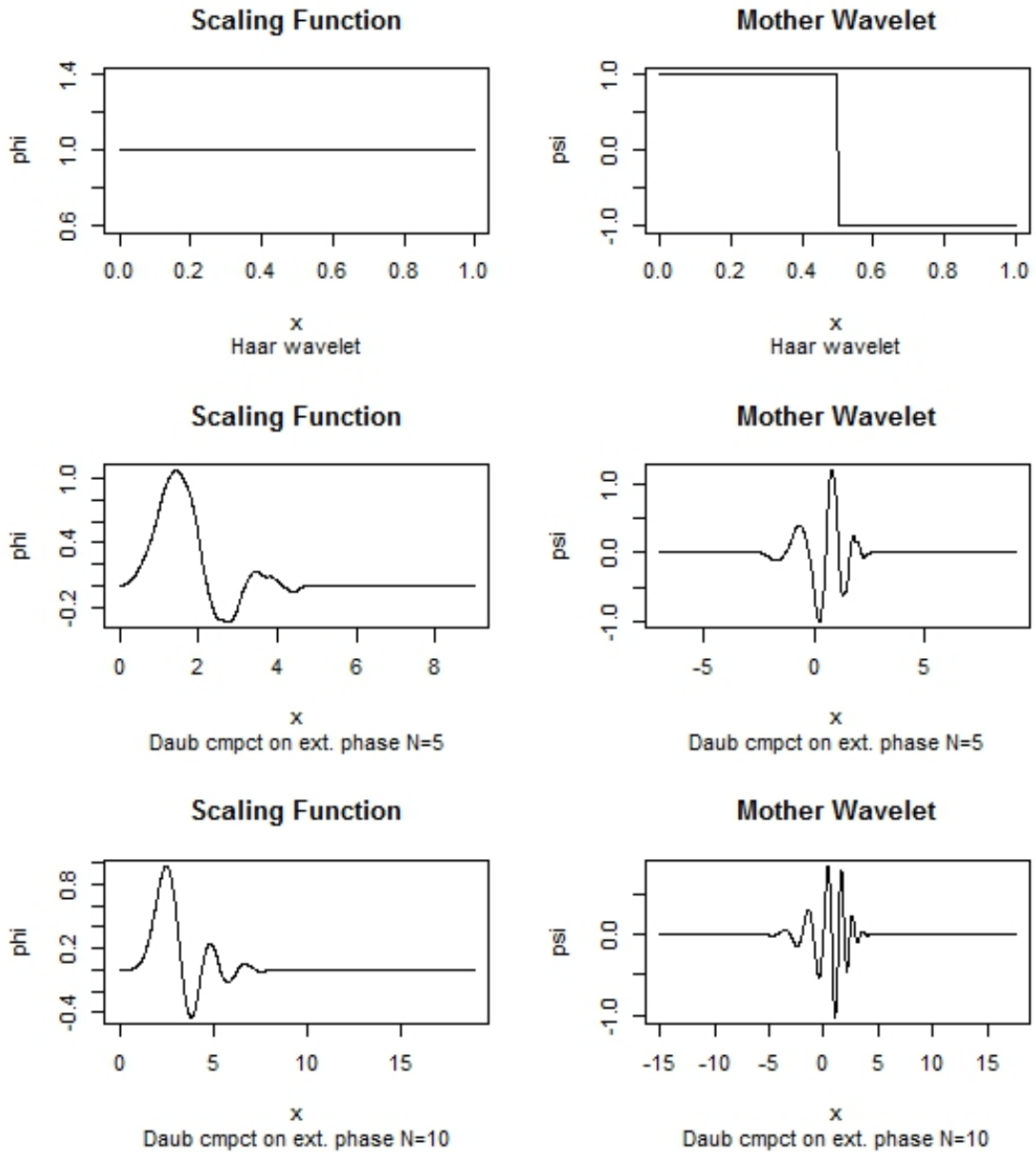

Figure 2.1: Left: Scaling Function Right: Mother Wavelet Top: Haar wavelet: $m = 1$; Middle: Daubechies Extremal Phase wavelet with $m = 5$; Bottom: Daubechies Extremal Phase wavelet with $m = 10$;

2.6.7 Haar in two dimensions

In this section we will briefly outline the two-dimensional Haar wavelet transform following Nason (2008) page 76 and Nason and Silverman (1994). This transform is due to Mallat (1989). The $2D$ wavelet transform applies to square matrices $\mathbb{X}n \times n$. Data types that can be represented in such a way are images and spatial data. It is also assumed that the $n \times n$ grid is regular and dyadic. The transform consists of the following three steps:

1. The filters $\mathcal{D}_0\mathcal{H}$ and $\mathcal{D}_0\mathcal{G}$ from equations (2.66) and (2.67) are applied to the n rows of $\mathbb{X}n \times n$ resulting in two $n \times n/2$ matrices denoted H and G

2. $\mathcal{D}_0\mathcal{H}$ and $\mathcal{D}_0\mathcal{G}$ are applied to the columns of H and G resulting in four $n/2 \times n/2$ matrices denoted HH, GH, HG and GG. The last three are stored as horizontal, vertical and diagonal finest level wavelet coefficients respectively.

3. $\mathcal{D}_0\mathcal{H}$ and $\mathcal{D}_0\mathcal{G}$ are applied to HH, then step 2 is repeated over the resulting HHH and GHH matrices yielding four $n/4 \times n/4$ matrices. Then this current step is applied to $HHHH$ recursively until there is one scaling coefficient left.

The generic scheme of the transform 2.6.7 is shown in Fig. 2.2. We will be using the two-dimensional Haar wavelet transform for spatial white noise and autocorrelation testing in chapter 6 of the thesis.

2.6.8 Wavelets in statistics

Nonparametric regression

During the past two decades, wavelets have found different uses in statistics, the most prominent being nonparametric regression estimation and time series analysis. This probably has to do with the fact that in both of those areas, we are dealing with function estimation from an imperfect realisation i.e. with *noise*. Furthermore, for different classes of problems, assumptions about the type of function such as smoothness can be made. In

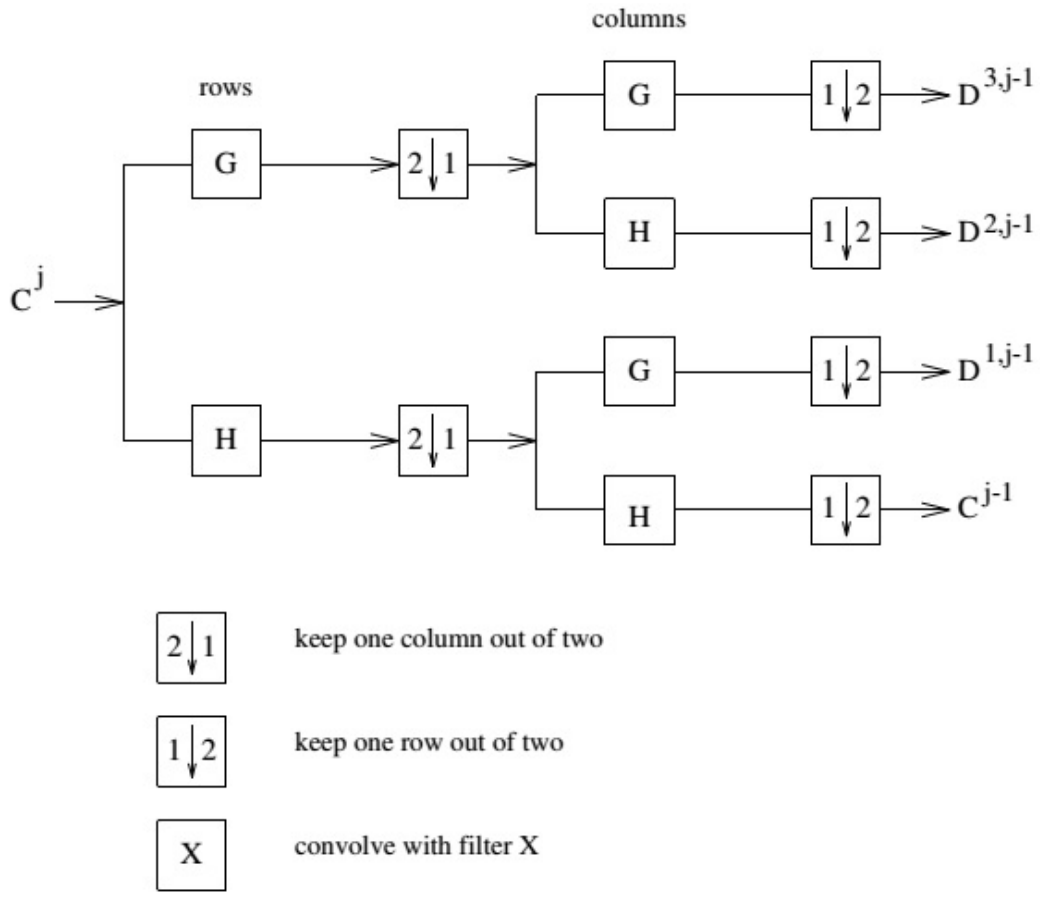

Figure 2.2: Illustration of generic steps 1 and 2 from the two-dimensional discrete wavelet transform algorithm. $C^j := \mathbb{X}_{n\times n}$ and D^1, D^2 and D^3 are the horizontal, vertical and diagonal coefficients resepectively. Reproduced with permission from Nason (2008).

this section we follow Nason (2008). The common estimation problem for nonparametric regression is:

$$y_t = f(x_t) + e_t, \quad t = 1 \dots T, \quad x_t = t/T, \tag{2.69}$$

where the noise e_t is often i.i.d. Gaussian i.e. $e_t \sim N(0, \sigma^2)$. As we have seen in the previous section, due to vanishing moments property, wavelets are good at representing low order polynomials. In particular, since wavelet coefficients of a smooth function with discontinuities can be sparse we can have an easy wavelet representation to work with. Then, because of Parseval's equality, we can apply techniques to the wavelet representation, which preserve the result to the function estimation.

One of the primary wavelet approaches for nonparametric regression is called wavelet shrinkage which boils down to *thresholding* the noisy wavelet coefficients rather than the observed data. It is due to Donoho (1993b), Donoho and Johnstone (1994), Donoho et al. (1995). The basic principle is the following: the discrete wavelet transform is applied to both sides of equation (2.69) and then some of the resulting wavelet coefficients are *thresholded* to zero and the inverse DWT is executed which gives us back the estimate of the function according to the data. Because of the sparsity of the DWT, it might be that the energy of the function is spread out to only a few wavelet coefficients. Because of the orthogonallity of the DWT, the wavelet coefficients of the noise e_t would again be noise, therefore the noise would be distributed across all wavelet coefficients (Nason, 2008) page84. Similarly to kernel density estimation, where the bandwidth is the crucial parameter, in wavelet thresholding there are different options for the *threshold* value. For example, we might decide on an absolute quantity and all wavelets less than it, to be equated to zero, so-called universal (hard) thresholding. Donoho and Johnstone (1994) developed a universal threshold equal to $\sigma\sqrt{2\log T}$. However, depending on the problem there are other options for more flexible thresholding such as soft thresholding — $\eta_\kappa = sgn(d)(|d| - \kappa)$, where κ is the threshold and d is a wavelet coefficient — adjusting the wavelet coefficients relative to their difference with the universal threshold.

2.6.9 Spectral estimation with wavelets

We will briefly outline two papers that deal with spectral estimation by wavelet thresholding. They are related to our work on white noise testing in chapter 4 in the sense that they are trying to estimate the actual spectrum of a time series, whereas we would be trying to find a departure from the white noise flat spectrum and gauge if there is enough evidence to reject the null hypothesis of white noise. Thus, we would be looking for wavelet coefficients of too large magnitude, based on a test for their distribution. On the other hand, spectrum estimation with wavelets basically uses those wavelet coefficients above the threshold in order to gauge the shape of the spectrum.

Donoho (1993a)

Section 4 of Donoho (1993a) shows an example of using the variance-stabilizing transformation of Wahba (1980) to the $\log$ of the periodogram yielding a so-called "Log-o-Gram", evaluated at the Fourier frequencies, of an AR(6) process. Then soft wavelet thresholding is applied to the "Log-o-Gram" which results in an estimate of the desired spectrum.

Thresholding of a tapered log-periodogram

Walden et al. (1998) refines the work of Donoho (1993a), by using a multitapered estimator of the log-periodogram. It is also explained that this helps with making the smoothed log-periodogram closer to the Gaussian distribution. In the paper different options for thresholding are explored — some depending on the scale of the wavelet coefficients. Several simulations are shown which confirm the suitability of the technique for spectrum estimation and different high-order AR and MA processes are considered. In another paper, McCoy et al. (1998) deal with the spectral estimation of so-called power law processes — having a spectrum in the form $f^{-\beta}$, where β is a positive exponent — by tapering.

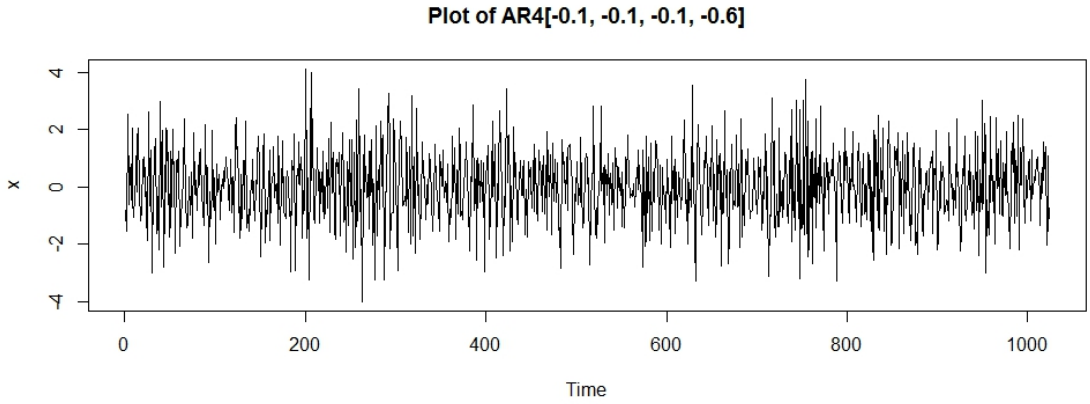

Figure 2.3: Realization of the AR(4) process from equation(2.70).

Vanishing moments in practice

Here we will apply the discrete wavelet transform with Daubechies extremal phase wavelets with five vanishing moments to an AR(4) process and its periodogram in order to illustrate why they are useful. In chapter 4, we will be using these wavelet coefficients to construct a test for a "flat spectrum". Let us have an AR(4) process defined by:

$$X_t = \rho_1 X_{t-1} + \rho_2 X_{t-2} + \rho_3 X_{t-3} + \rho_4 X_{t-4} + \varepsilon_t, \quad t = 1, \ldots, 1024 \tag{2.70}$$

where $\rho_1 = \rho_2 = \rho_3 = -0.1$; $\rho_4 = -0.6$ and ε_t is Gaussian white noise.

Due to its negative parameters, the process from equation (2.70) has a high oscillation rate, in other words its spectrum is dominated by high frequencies, although due to the small magnitude of the first three parameters, we could also expect some energy at the lower frequencies. Its spectral density has two peaks. Fig. 2.3 shows the realization of the AR(4) process. Fig. 2.4 shows the periodogram and its wavelet coeffcients. Although erratic, the periodogram shows the two peaks — at lower and higher frequencies. The few large wavelet coefficients reflect well the peaks of the periodogram and most of the rest are very close to zero. Hence our assertions from the previous sections are confirmed and when doing a statistical test of the wavelet coefficients we would expect that the large coefficients' p-values to be small. We will explore these assertions in chapter 4.

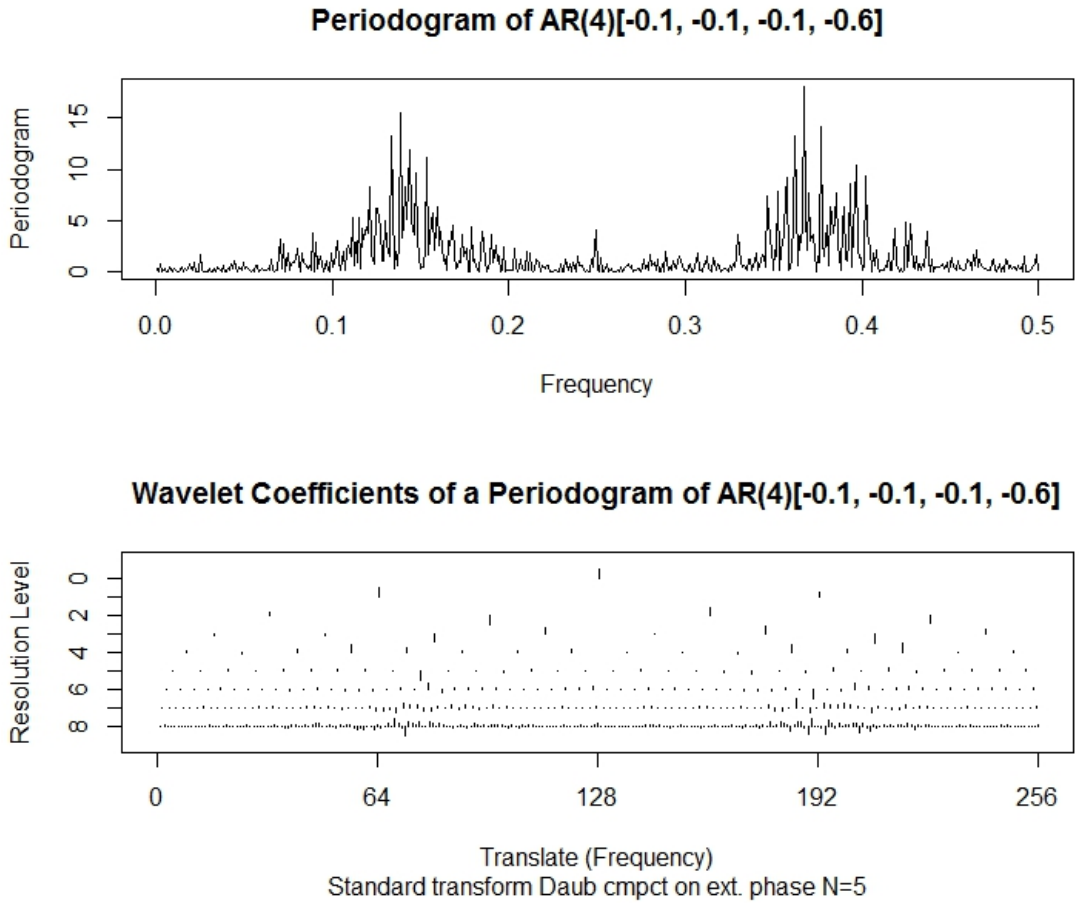

Figure 2.4: Top: The raw periodogram of the AR(4) process from equation(2.70); Bottom:Its wavelet coefficients of raw periodogram from a Daubechies extremal phase wavelet with 5 vanishing moments.

Conclusion and directions

Both techniques used for applying wavelet thresholding to the (log)periodogram use some type of transformation that would help its distribution to be close to Gaussian. Smoothing by tapering, on the other hand, helps in representing the periodogram closer to a polynomial. Then wavelet thresholding achieves very good results because of wavelets decorrelation property as well as vanishing moments. For a detailed source of wavelet methods for signal analysis and thresholding methods in both time and frequency domains, please consult Percival and Walden (2000) chapters 9 and 10.

In this thesis we will use an approximation to the distribution of the wavelet coefficients of the raw periodogram in order to be able to discern whether or not the series is (Gaussian) white noise. However, the mentioned properties of wavelets help us detect non-whiteness even when the series are not Gaussian — shown in our simulation study in chapter 4.

Chapter 3

Literature Review — Functional Data Analysis

3.1 Overview of Functional Data Analysis

Functional Data Analysis (FDA) is a branch of statistics that deals with data arising as curves or surfaces. They can arise when we observe a continuous process via a discrete grid of measurements. The term was coined in a seminal paper by Ramsay and Dalzell (1991). Data as curves occur in many natural phenomena. Examples include spectrometric analysis and electricity load data as in Ferraty and Vieu (2006), temperatures from weather stations and human growth curves as in Ramsay and Silverman (2002) and magnetometer curves in Hörmann and Kokoszka (2012). On the other hand, representing continuous time stochastic processes as sequences of random variables in function spaces is an established tool in probability and statistics, see Doob (1953) or Grenander (1981).

Mathematically, the functional paradigm means that we are dealing with functional space such as the Hilbert L_2 space of integrable functions, rather than standard Euclidian n-dimensional space ($\mathbb{R}^n$) i.e. our observations are entire functions and not points. Statistics, in this domain, leads to theoretical challenges requiring more sophisticated arguments compared to those found in multivariate analysis. For instance, when we know

that our process produces data in the form of smooth curves, then we need an infinite-dimensional vector space of functions which accounts for that e.g. Sobolev spaces and/or regularization. Moreover, this means that many functional data must be pre-processed before analysis. For example, pre-processing is necessary because we know that the underlying function is either smooth or periodic, thus requiring an appropriate basis expansion in order to emphasize those characterizing features.

Furthermore, we need a way to measure the dynamics of the curves within the space i.e. the need for a different (semi)norm and (semi)metric than the Euclidian one. This comes from the fact that we are interested in the difference or similarities in functions' dynamics and not just in the distances between them. Moreover, when we think about time series — consider how many curves might be on an interval - there are infinitely many of them which necessitates the use of infinite-dimensional spaces. Also, if we wish to predict or classify functions, we need a specialized metric to measure similarity of the curves, leading to a way to construct an estimator in the regression sense or a kind of nearest neighbour distance for clustering.

3.1.1 An early FDA problem

One of the first research articles dealing with the functional data problem is Besse and Ramsay (1986). It poses for the first time the functional data problem, illustrated with real data from human tongue dorsum movement.

The curves in Fig. 3.1 follow closely a linear differential equation. In order to accommodate this in the paper, the authors use information from a derivative metric in order to model the data. The authors show that this is equivalent to a change of metric in the row space of classical principal component analysis(PCA). Another important feature of their article is that it shows how the standard least-squares regression is not very useful to model such type of data because it cannot verify that the data come from regular functions — the polynomial splines which have been used to pre-process the data.

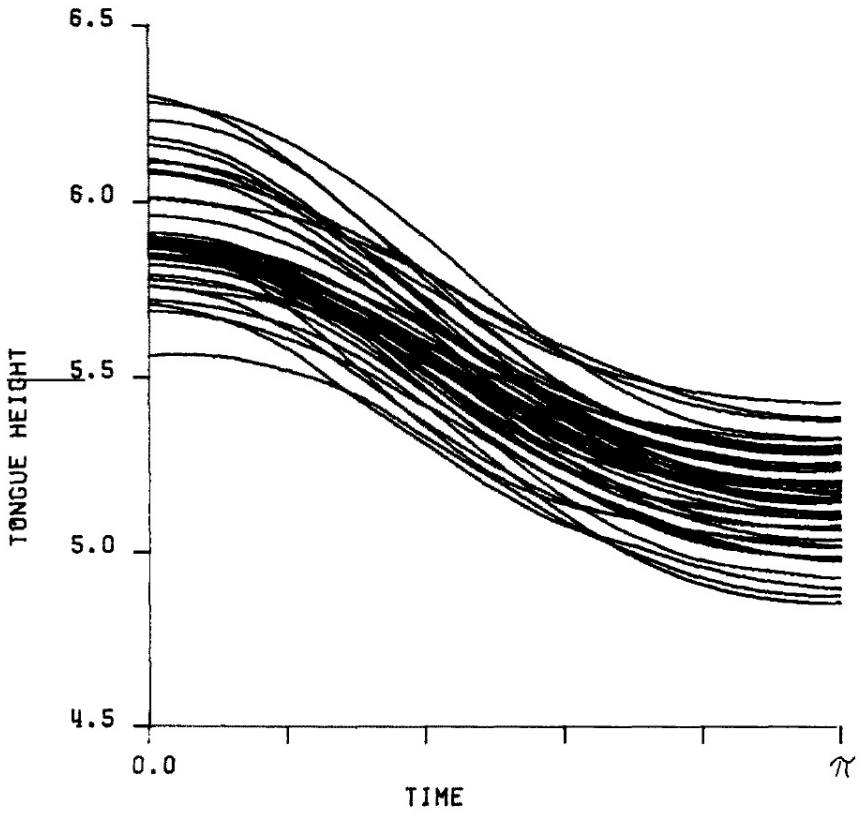

Figure 3.1: These curves represent the tongue dorsum height while pronouncing the sound 'Kah'. They are result of polynomial spline smoothing of tongue position sampled every milisecond using an ultrasound sensing technique. Each record begins and ends at the point where the slope is zero. Here the lengths of the curves have been standardized to the interval $(0, \pi)$. Picture reproduced with permission from Besse and Ramsay(1986)

3.1.2 Tools for Analysis

In order to adapt to the nature of the curves, Besse and Ramsay (1986) make the assumption that the sampled functions lie in a vector space H^m of functions with $m-1$ continuous derivatives i.e. a Sobolev space, defined on an interval $T = [a, b]$. It is also assumed that the m-th derivative has a finite Lebesgue integral over T. This means that lower-order derivatives are differentiable. This setup leads to the idea that a function $T \to R$ in this space might be represented as a signal function + noise function, where the signal function satisfies a homogeneous linear differential equation. There are also some boundary constaints which are not discussed here. An important feature is the usage of reproducing kernel for these Hilbert spaces (RKHS). RKHS for a Hilbert space of functions H, defined on an interval T, is a bivariate function $k(.,.)$ on T$\otimes T$, called kernel, which satisfies the *reproducing equation*:

$$< k(s,.), x >= x(s), \qquad x \in H \tag{3.1}$$

where $< .,. >$ is the dot product from Def. 3.2, the point in $k(s,.)$ resembles a variable that is integrated out in the dot product. The RKHS are closely related to Green's functions with respect to a differential operator. The idea of using RKHS in the concrete examples with the tongue curves is that it will help dividing the space into a signal and noise subspaces. This also supports the idea that the basis representation used for preprocessing sampled functions is important with respect to their subsequent modelling. The main finding of their paper is that reproducing kernels help in the formulation of the spline interpolation problem and RKHSs help developing functional Principal components analysis (fPCA) which is a cornerstone of functional data analysis. We will talk more about fPCA in the next section.

Besse and Ramsay (1986) led to the crucial question what are the different metrics and basis representations possible in the functional context. The greater freedom in basis

and metric choice is good, but adds to complexity and scalar models cannot always be straightforwardly extended to the functional context.

3.2 Functional Data Analysis Framework

In this section we will describe the main pillars on which FDA is founded.

3.2.1 Functional Space, Mean and Covariance Functions

As already mentioned in the previous section, functional data are curves $\mathcal{X}(t) \in H$ where H is usually some Hilbert space with certain properties and $t \in T$, where $T = [a, b]$ is an interval. In this thesis, H will be the $L^2_{[a,b]}$ space of square-integrable functions on the interval $[a, b]$. The standard setup of H is a separable space with dot product inducing the norm:

Definition 6. *Dot product in H. Let the dot product of $x, y \in H$ be:*

$$< x, y >= \int_a^b x(t)y(t)dt \qquad x, y \in H \tag{3.2}$$

An important statistical consequence, from the change of paradigm, is the fact that the mean value is no longer a scalar, but a function itself. Let us define our functional random variable (f.r.v.) $\mathcal{X} \in H$. Let also $T = (0, 1)$. Then the observation of our f.r.v. is $X(t)$ and the mean function is:

$$E\{X(t)\} = \mu(t). \tag{3.3}$$

A similar adaption holds for the covariance/correlation which become operators. Loosely speaking, an operator is a mathematical object that takes a function and then produces another function. In Hilbert spaces this usually happens through dot product, denoted by

$<,>$ i.e. integration. For example the covariance operator of a zero-mean random variable $\mathcal{X}$, taking values in a Hilbert space H is:

$$C_{\mathcal{X}}(x) = E(<\mathcal{X}, x>\mathcal{X}), \qquad x \in H. \tag{3.4}$$

However, as we will always work with discrete data, we would use the sample covariance function. Suppose data is zero-mean, then

$$C_{X,Y} = E\{X(s)Y(t)\}, \qquad s,t \in T. \tag{3.5}$$

In the FDA literature, $C_{X,Y}$ is called the covariance kernel. Thus, when one wants to 'look at the correlations', there are different possibilities. For example, if we had just two functional variables on a grid of size 100 points i.e. we have 100 values from 0 to 1 for t and s respectively, this means that there are almost 5000 correlation pairs to look at, which can be unmanageable. Therefore, one must think of new ways to represent data and relationships when they are functional.

A possible solution could be a contour plot of the correlation functions as in Ramsay and Dalzell (1991)which builds on Besse and Ramsay (1986). More precisely, they show how to build generalizations of the linear model in functional terms with the help of L-splines and how the Green function links RKHS with the theory of ordinary differential equations. Their methodology is applied to Canadian weather stations data. Moreover, the authors clearly distinguish three main stages in functional data analysis:

1. Choice of function space in which the analysis is about to take place.

2. Specification of the analysis in functional analytic terms.

3. Determination of how a finite-dimensional observation vector is to be mapped into a function space. For instance, for the regular curves as in Fig. 3.1, this step could be realized through the use of smoothing splines.

Furthermore, the mean and variance functions themselves, being smooth curves the-

oretically, would require some smoothing procedure. The first article that poses that problem and shows an extended solution of the classical second derivative regularization/penalization procedure is Rice and Silverman (1991).

3.2.2 The Ramsay and Silverman monograph.

The case for independent and identically distributed functional data is explored in a monograph that deals with most of the conceptual, theoretical and practical issues as well as the translation of classical parametric methods for functional data — Ramsay and Silverman (1997, 2002). The book describes ongoing research and discusses different ways to generalize procedures to function spaces. It starts with exploring techniques for the displaying of functional data in the first two chapters, such as pairwise plot of derivatives and some tools such as contour plots and phase-plane plots. Chapter 3 explores the major mathematical orthonormal bases that could be used to represent functions such as Fourier, splines, B-splines, wavelets and polynomial — to name a few. Only then, the authors start with generalizing least squares methodology to the functional case and regularization procedures, in chapters 4 and 5. Chapter 7 describes the techniques to display and recording of functional data as this is a challenge in its own right. The rest of the book, apart from the formulation of the general functional linear model, contains several chapters devoted to the generalization of principal components analysis — such as principal differential analysis which uses differential operator instead of projection operator for defining the most important variance basis directions. There are also chapters on the functional generalizations of canonical correlation and dicriminant analysis. Most of the necessary theory, for example Green's functions and RKHS, is provided in the appendicies of the book.

An important point is also made in several chapters — when thinking about statistical modelling of functional, multivariate and scalar data together, there are mixed cases. For example, we might have a functional predictor and a scalar outcome, or even, a multivariate dataset in which we have both types of predictors towards a functional or scalar outcome. When thinking about functional data, we need a basis representation for them,

but for the scalar data we do not. However, when we want to have a unified model for such a mixed dataset there are different trade-offs to be made in order to balance practical significance with theoretical foundations.

3.2.3 Nonparametric FDA: Ferraty and Vieu (2006)

The nonparametric paradigm for functional data is developed both theoretically and with many real data examples in Ferraty and Vieu (2006). This book discusses practical ideas in the first two chapters to set the scene. Then chapter 3 discusses the problem of an appropriate space for functional data as well as closeness notions for functions such as (semi)metric. In chapters 4, 5 and 7 the kernel estimation methodology is generalized for functional data as well as some classical procedures such as conditional: mean, median and mode estimation and quantile estimation. Chapter 6 develops asymptotic theory and introduces a rarely used concept such as almost complete convergence for infinite-dimensional processes. The standard convergences in probability and almost surely are special cases of this almost complete convergence. There are also separate chapters dealing with nonparametric supervised and unsupervised classification procedures such as k-nearest neighbours algorithms and heterogeneity indicies. Moreover, two theoretical chapters — chapter 10 dealing with mixing conditions and chapter 11 with asymptotics for dependent data — constitute a separate part of the book. The book contains many real data examples and there is a companion website, hosting R scripts of most of the computational procedures.

3.3 Some key tools in Functional Data Analysis

This section reflects briefly on the principal established tools in FDA: how they relate to each other and how they can be combined.

3.3.1 Functional Principal Components Analysis

Possibly the most common tool for dealing with functional data is the functional Principal Components Analysis (fPCA). The method has deep foundations in what is called Karhunen-Loeve expansion of functions which is the functional generalization of the matrix eigendecomposition. The difference between standard and functional PCA is that for the functional, some smoothing of the results is usually performed. This is required because the eigenfunctions are assumed to be smooth and also, as we do not observe the process continuously, some noise is present in the empirical eigenvectors. The smoothing can bea achieved either as a pre-processing step directly on the data or as a a posterior regularization step to the standard principal components derived. The smoothing is repeatedly shown in Ramsay and Silverman (2002). As with any basis problem, many mathematical and orthonormal bases are possible. However, Ramsay and Silverman (2002) demonstrate that good basis selection is highly problem-dependent.

Another distinguishing feature of FDA, compared to standard multivariate statistical analysis, is the possible usage of the fPCA. In standard multivariate context, the PCA is often an exploratory tool. However in functional data analysis, it can be used directly for modelling as noted in Ferraty and Vieu (2006). This happens through the use of a semi-metric, which we will explicitly define when we look at local weighting of functional variables in section 3.3.4.

3.3.2 The Need for Different Norms and Metrics

Ramsay and Silverman (2002) explore different kinds of norms and metrics, for instance those including derivatives. Derivatives arise naturally, because when we want to model functions, we are not simply interested in the Euclidian distances between them. The dynamics of the functions themselves are often far more important if we want to classify or forecast them. As we know from fundamentals, the velocity of a particle is the derivative of its position vector with respect to time and the acceleration is the second derivative.

This observation has a direct application for some functional datasets and can be used to classify curves or to reveal interesting relationships among the different functional observations. A famous example in the FDA literature is the children's gait data, analyzed in Ramsay and Silverman (2002). For example, simple plotting of the estimated first derivatives together with the average derivative could show anomalous records. Furthermore, plotting first versus second derivatives could tell us much about the motion characteristics of the subjects.

3.3.3 Combining of fPCA and derivative metrics

In the example of Fig. 3.1, Besse and Ramsay (1986) suggest the curves might be modelled by a linear differential operator. This technique is matematically similar to standard principal components derived from a multivariate covariance matrix. The essential operatrion in PCA is projection on the eigenvectors of the covariance matrix. This can be expressed with a projection operator. In Ramsay (1996); Ramsay and Silverman (2002) the authors develop the concept called Principal Differential Analysis (PDA) which is similar to PCA, but acts with respect to a linear differential operator. Another famous dataset analyzed in Ramsay and Silverman (2002) is the Canadian weather stations data. They are analyzed by the means of a linear differential operator. The initial hypothesis is that the variation should be mainly sinusoidal, however it turns out that for continental weather stations, there are strong spring and autumn systematic effects due to "change in reflectance of the land". Ramsay and Silverman (2002) explain that for those stations a non-homogeneous differential equation could be a better model.

3.3.4 Other Aspects of Functional Data Analysis

In this section we touch on a few features of the already mentioned semimetrics. The section follows Ferraty and Vieu (2006).

Semi-metrics and PCA

As mentioned earlier, a crucial task in FDA is the possibility of using different norm from the Euclidian one. Following Ferraty and Vieu (2006) chapters 3 and 4, let us start with the definition of a semi-metric first.

Definition 7. *d is a semi-metric on a normed space H iff*

$\forall x \in H, \qquad d(x,x) = 0$

and

$\forall (x,y,z) \in H \otimes H \otimes H, \qquad d(x,y) \leq d(x,z) + d(y,z)$

In fact, the only condition that distinguishes the semi-metric from the metric is that:

$$d(x,y) = 0 \Rightarrow x = y \tag{3.6}$$

is not true for a semi-metric i.e. if the first derivatives of two functions are the same, it does not mean that the functions themselves are the same.

Let $H = L^2_{[0,1]}$ be the Hilbert space of square integrable functions on the interval $[0,1]$ and let $\mathcal{X} \in H$ Also let $T = (0,1)$ and $t \in T$.

An interesting semi-metric for functional data can be built from functional PCA which arises from the Karhunen-Loeve expansion of the functional random variable $\mathcal{X}$ by writing.

$$\mathcal{X} = \sum_{k=1}^{\infty} \left(\int \mathcal{X}(t) v_k(t) dt \right) v_k \tag{3.7}$$

where $v_1, v_2, ..$ are the orthonormal eigenfunctions of the covariance operator of the realization of $\mathcal{X}$.

In terms of data, however, $\mathcal{X}$ is represented by a matrix X of dimensions n by p, with n being the number of curves and p being the number of discretization points. Thus, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, $t = j/p$ This expansion can be truncated at, say, the q^{th}

eigenvector which leads to the family of semi-metrics:

$$d_q^{PCA}(\mathcal{X}_i, \chi) = \sqrt{\sum_{k=1}^{q} \Big(\int [\mathcal{X}_i(t) - \chi(t)] v_k(t) dt \Big)^2} \tag{3.8}$$

Of course there is no a direct way to compute the above integral as we observe the functions discretely. That is why the semi-metric must be estimated using some sort of weights, depending on the grid or quadrature rules. Also, it is possible to use a semi-metric within the PCA itself which would reveal features of the dataset.

3.3.5 Functional time series: Bosq(2000)

Bosq (2000) develops the theory of functional time series and deals predominantly with Autoregressive Hilbertian Processes(ARH), defined next in section 3.4. It is a book that develops mathematically the ARH process and also considers its extensions to Banach spaces. Chapters 1 and 2 lay out the necessary probability theory for covariance operators, random variables and their sequences in Hilbert and Banach spaces. Chapter 3 is devoted to the ARH(1) process model and its existence and limit theorems — we use Theorem 3.6 from this chapter in order to develop our ARH(1) order verification procedure in chapter 6 of this thesis. Chapter 4 deals with the theoretical estimation of the ARH(1) from representation in a countable basis, the eigenelements of the covariance operator and their convergence in distribution. Chapter 5 extends the theory to ARH(p) and chapter 6 does the extension from Hilbert to Banach spaces. Chapters 7 deals with general linear processes in function spaces and their existence and invertibility conditions, while chapter 8 gives theoretical guidance of how the autocorrelation operator of the process could be estimated within the space $C[0, 1]$ and used for prediction. Finally, chapter 9 gives some applied references and methodological guidances for estimation such as generalised cross-validation (CV) procedure for the covariance operator.

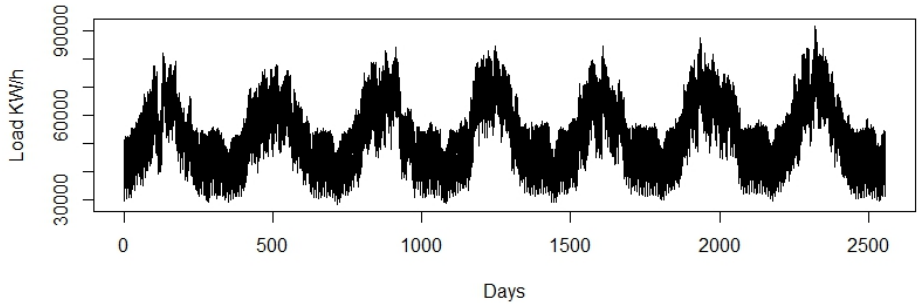

Figure 3.2: EDF daily electricity load time series from September 2002 till August 2009

3.4 Functional Time Series

FDA comes up as a natural framework when we consider time series processes that are observed on consecutive intervals. These include transaction curves, electricity load curves, wind speed curves. When we observe and measure such a process, it will often have a seasonal component with period equal to the interval and may be repetitive over time. For example, electricity load that is measured every half-hour producing 48 measurements every day. Example daily data are plotted on Fig. 3.3. Fig. 3.2 shows daily data for 7 years and there is inherent monthly and quarterly periodicity. Those data have been provided to us by Xavier Brossat from Électricité de France. Depending on the representation, there is also weekly periodicity not shown here. So, we could choose our interval to be one day and our function is a curve throughout that day or we could decide to smooth it weekly — many options are possible. Furthermore, we might even have irregularly spaced data and then we could use a spline-smoother and thus to come up with regular design, resulting in one smooth curve per day. In this section we will mainly follow Bosq (2000).

3.4.1 The Autoregressive Hilbertian Process of Order One

One of the seminal articles in functional time series is Bosq (1991). It formulates for the first time the functional autoregressive model of order one, also called Autoregressive Hilbertian process of order one (ARH(1)) which is a functional generalization of the clas-

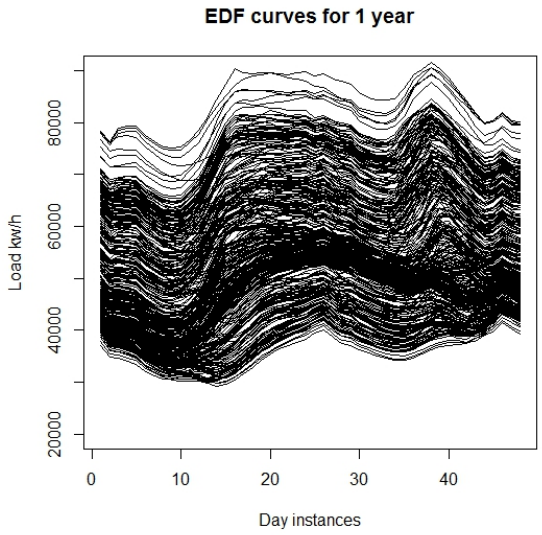

Figure 3.3: This picture shows the Electricite de France electricity load curves from 1 Sep 2008 to 31 Aug 2009. On the x axis we have the 48 instances of daily measurement and on the y is the load in kw/h

sical autoregressive process of order one. The process is defined to be stationary similarly to the scalar and multivariate AR(1) process cases. However, an interesting practical difference is that when one models functional data, trends/features are not always removed, because they are contained in all curves, thus contribute to the modelling. The theory for the order one process and for higher orders, as well as processes in Banach spaces is developed in Bosq (2000). In the subsequent sections we will define the ARH(1) and the sampling mechanism that is inherent with such continuous-time processes as well as state some open problems.

3.4.2 Notation and Theoretical Setup for ARH(1)

Following Bosq (2000), functional time series are curves defined as: $\{\mathcal{X}_n, n \in \mathbb{Z}\}$ where each $\mathcal{X}_n$ is a function $\mathcal{X}_n(t), t \in [a, b]$. As described, those curves reside in the Hilbert space $H = L^2_{[a,b]}$ of square-integrable functions on the interval $T = [a, b]$. For the exposition in this thesis, the interval T will be normalized to $[0, 1]$ and thus each realization will be defined on this interval. Then each $\mathcal{X}_n(t) \in H$, $t \in T$ is a function in H. This paradigm may be represented by associating a sequence of random variables i.e. time

series $\xi = (\xi_t, t \in \mathbb{R})$ taking values in the function space H. This is obtained by setting:

$$\mathcal{X}_n(t) = \xi_{nh+t}, \qquad 0 \leq t \leq h, \qquad n \in \mathbb{Z}, \quad t \in T \tag{3.9}$$

Thus, $\{\mathcal{X}_n, n \in \mathbb{Z}\}$ is infinite-dimensional discrete-time process. In our case h would be 1 and t will correspond to the grid on which the function is evaluated in $[0, 1]$. This is useful when the data contain a seasonal component of length h, for instance, when we observe daily realizations of the process

Let us now lay down a set of definitions, necessary to define the ARH(1) process.

Definition 8. *Compact Operator on a Hilbert Space H.*
An operator l, on a Hilbert Space H is compact if there exists two orthonormal bases of H, (e_j) and (f_j), and a sequence (λ_j) of real numbers such that:

$$l(x) = \sum_{j=1}^{\infty} \lambda_j < x, e_j > f_j, \quad x \in H \quad s_1, s_2 \in \mathcal{S} \tag{3.10}$$

Equation (3.10) is called the spectral decomposition of l

Definition 9. *Hilbert-Schmidt operator*
A compact operator l is Hilbert-Schmidt if:

$$\sum_{j=1}^{\infty} \lambda_j^2 < \infty \tag{3.11}$$

where λ_j is from Def.8

Definition 10. *The Space of Hilbert-Schmidt Operators*
The space $\mathcal{S}$ of Hilbert-Schmidt Operators is a separable Hilbert space with respect to the scalar product:

$$< s_1, s_2 >_{\mathcal{S}} = \sum_{1 \leq i,j \leq \infty} < s_1(g_i), h_j >< s_2(g_i), h_j > \tag{3.12}$$

where (g_i) and (h_j) are two arbitrary orthonormal bases of H and s_1, s_2 are Hilbert-Schmidt operators.

Definition 11. *The Hilbert-Schmidt norm on a space $\mathcal{S}$*

$$\|s\|_{\mathcal{S}} = \left(\sum_{j=1}^{\infty} \lambda_j^2\right)^{1/2} = \left(\sum_{j=1}^{\infty} s(g_i)^2\right)^{1/2}, \quad s \in \mathcal{S} \tag{3.13}$$

where λ_j is from Def.8 and (g_i) is an arbitrary basis of H

Definition 12. *Weak and strong orthogonality*

H-valued random variables X and Y are called:

- weakly orthogonal *if* $\mathrm{E} < X, Y >= 0$
- strongly orthogonal *if* $C_{X,Y} = 0$

Stochastic independence implies strong orthogonality, which implies weak orthogonality, but not in reverse.

Definition 13. *Functional white noise*

A functional or H-valued strong *white noise ε_n is a sequence of zero-mean, independent and identically distributed random variables taking values in the Hilbert space L_H^2. If they are not mutually independent, but* strongly orthogonal*, they are called just H-valued white noise.*

Finally, The ARH(1) is defined in the following way:

Definition 14. *$\mathcal{X}_n$ is an ARH(1) process if and only if*

$$\mathcal{X}_n = \rho(\mathcal{X}_{n-1}) + \varepsilon_n \qquad n \in \mathbb{N} \tag{3.14}$$

where ρ is an infinite-dimensional linear autocorrelation operator i.e. $\rho = C_{\mathcal{X}_1} C_{\mathcal{X}}^{-1}$ where $C_{\mathcal{X}}(x) = E(< \mathcal{X}, x > \mathcal{X})$ is the covariance operator, $C_{\mathcal{X}_1}(x) = E(< \mathcal{X}_n, x > \mathcal{X}_{n+1})$ is the lag one covariance operator ε_n is a functional H-valued white noise, for instance a Brownian bridge or a Wiener process.

The process (3.14) is defined to be stationary in a similar way to the scalar and multivariate autoregressive process of order one cases. The condition which ensures convergence in the MA(∞) representation, similar to the scalar AR(1) case, is $\|\rho^{j_0}\|_{\mathcal{S}} < 1$ where j_0 is an integer and $\mathcal{S}$ denotes the norm in the space of Hilbert-Schmidt operators as in Def.11. It should be noted also that the autoregression in (3.14) could be more general with a nonlinear operator, for instance, ρ could be estimated nonparametrically via a functionally adapted kernel estimator as in (Ferraty and Vieu, 2006).

3.4.3 The Data Problem in Functional Time Series

Although we use continuous models such as (3.14), in practice they are presented to us, and evaluated on, a discrete set of values e.g. points over a grid that represents the interval. Fig.3.4 shows a simulated realization of ARH(1) using the `far` package in `R` from Damon and Guillas (2010). This means that, in practical terms, functional data analysis is similar to standard multivariate analysis, but the underlying model is different. Hence, our functional data will be respresented as an $n \times p$ matrix where n is the number of discretized curves and p is the number of points on the grid. Theoretically, the number of gridpoints p goes to infinity to completely capture a curve. This means that the finer the resolution, the better the representation. However, another practical problem is that when we have a measurement instrument, say magnetometer, and we use it continuously its data inevitably contain some measurement error. That is why, one way to model functional data is to smooth the eigenvectors of their covariance matrix or to truncate the covariance eigendecomposition and use projections of the original data on this set of basis vectors, as considered in Ramsay and Silverman (2002). Hence, a key tool for building the functional linear model is to upgrade the general linear model. Another typical technique is to use some appropriate basis expansion to pre-process the data, such as spline or Fourier basis functions.

With respect to functional time series, the grid nature also has more implications. For instance, if we wanted to predict the future values of the function for a whole year and we

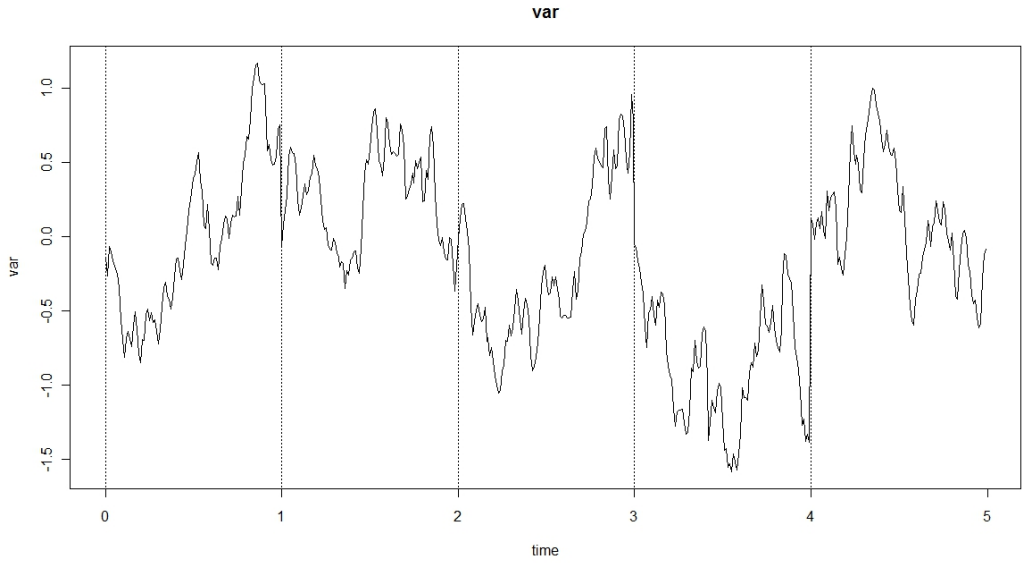

Figure 3.4: 5 curves of a Simulated Realization of ARH(1) with Wiener Noise over a grid of 100 points

had resolution of one observation per half hour, then it would not be wise to use all the gridpoints in the estimation, because this might incorporate a substantial amount of noise in the prediction. We could do some averaging or smoothing on the week level in order to predict months and thus the year. We could also do some differencing or centering operation on the data or even to model the different level of resolutions with various techniques — a good example is Cho et al. (2013). For example the weekly average curves could be subracted from all curves.

Moreover, since later we will be interested in predicting a trajectory, it would not be wise to use only the principal directions of variance, as in standard multivariate principal component analysis, either. This phenomenon was shown with simulated data in Hörmann and Kokoszka (2012). That is why the eigendecomposition of a covariance/correlation matrix (in a particular lag) must be used carefully with respect to prediction of functional data. A main trade-off in forecasting functional time series is how many principal components of the data to use for prediction? There exist different cross-validation procedures in order to do that as suggested in Bosq (2000) and Besse et al. (2000). The recommended

approach is to choose the number of principal components based on a validation set of the data and select the number of principal components which gives the least empirical prediction error. We will use this approach and suggest another in chapter 7.

3.4.4 Interesting problems in time series FDA

Time-varying ARH and/or local stationarity and functional spectrum

An interesting model is the time-varying functional AR process. This means that our autocorrelation operator changes over time. Let us have our data in a matrix form X_{np} where n is the number of curves and p is the number of discretization points. Then we could have the model $X_n = \rho_1(X_{n-1}) + \epsilon_n$ for $n \leq n_1$
and $X_n = \rho_2(X_{n-1}) + \epsilon_n$ for $n > n_1$ for some integer n i.e. if $n = 1000$, say $n_1 = 500$
Most of the present literature deals with either stationary models such as ARH (1) or use complicated curve clustering procedures, as in Antoniadis et al. (2013) in order to fit the stationary model to segments. Another possible direction, which has not been much explored is defining a functional spectrum and locally stationary versions of the theory.

Order Determination for ARH

Despite the popularity of the ARH model in the field, there appears to be little literature addressing its suitability for a functional time series dataset. An exception is the paper from Kokoszka and Reimherr (2013) dealing with this question. The ARH(p) model is similar to AR(p) in having operators of higher lag in definition 14. Kokoszka and Reimherr (2013) develop specific representation and estimation routines and a test statistic for the ARH(p) order determination through a functional linear model, suited for the testing problem of order determination. We intend to investigate an aspect of ARH order determination later in the thesis in chapter 7.

Chapter 4

Univariate Wavelet White Noise Tests

This chapter is based on our published paper Nason and Savchev (2014b).

Testing whether a time series is consistent with white noise is an important task within time series analysis and residual checking. We develop three new fast and efficient white noise tests by assessing spectral constancy via the wavelet coefficients of a periodogram. Our first Haar wavelet-based test derives the approximate distribution of the wavelet coefficients of the asymptotic periodogram for independent and identically distributed data under mild conditions. Our second test uses a single Haar wavelet coefficient obtaining a test statistic as a linear combination of odd-indexed autocorrelation values. To achieve greater power our third (general) test uses compactly supported Daubechies wavelets. We prove that the general wavelet test coefficients are asymptotically normal and derive a formula for its theoretical power for an arbitrary spectrum. We show examples for some autoregressive moving average models for various sample sizes that exhibit close theoretical agreement with simulation. We present a simulation study showing that our tests are broadly competitive, sometimes perform extremely well and exhibit them on a wind speed time series.

4.1 Introduction

Testing for white noise is a cornerstone in time series analysis. Such tests can be useful in their own right or part of a larger procedure that assesses model fit residuals for remaining stochastic structure. A process is white noise if and only if its spectrum is flat and if and only if the wavelet coefficients of the spectrum are all zero. Hence, our new tests works by statistically testing whether the wavelet coefficients of the periodogram are significantly different from zero.

When using Haar wavelets it turns out that we can establish the exact distribution of the coefficients from the asymptotic distribution of the normalized periodogram for independent and identically distributed data under mild conditions (the null). We also introduce a related test based on assessment of a single Haar wavelet coefficient that has a simple representation in terms of odd-indexed autocorrelations. To improve statistical power of our test we replace Haar wavelets by more general wavelets and show that, asymptotically, the coefficient distribution is Gaussian with a specified mean and variance. Further, we develop a theoretical formula for the power of the general wavelet test for fixed T for any spectral density function. The theoretical power function can give guidance on questions such as 'how large does my sample have to be to detect departures from white noise against a specific alternative'?

Why do we need another white noise test? All tests are 'directional' in that they have excellent power for some kinds of alternative, reasonable power for many alternatives and poor power for others. Our test has excellent power for some alternatives that other tests do not achieve and so, it is, we believe, a useful addition to the literature. Moreover, our test benefits from an implementation in R, theoretical backup and a theoretical power formula.

4.2 Building blocks of our tests

4.2.1 Basic Components

Suppose that $\{X_t\}_{t\in\mathbb{Z}}$ is a real-valued second-order stationary stochastic process with mean $\mu < \infty$, variance $\sigma_X^2 < \infty$, autocovariance $\gamma_X(k)$ for integers k and associated spectral density function $f(\omega)$ for $\omega \in [-\pi, \pi) = \Pi$. We address the problem of testing the hypothesis $H_0 : \{X_t\}$ is white noise versus the alternative H_A that $\{X_t\}$ is not white noise given a realization $X_t, t = 1, \ldots, T$ for some positive integer T.

The process $\{X_t\}$ is white noise if and only if its spectral density function $f(\omega)$ is a constant function on Π. The spectral density function, f, can be estimated from a realization $\{X_t\}_{t=1}^T$ via the periodogram

$$I_T(\omega) = (2\pi T)^{-1} \left| \sum_{t=1}^{T} X_t e^{-i\omega t} \right|^2, \tag{4.1}$$

which can be computed at the Fourier frequencies $I_p = I_T(\omega_p)$, where $\omega_p = 2\pi p T^{-1}$ for $p = 1, \ldots, T/2$.

Our tests will be based on wavelet decompositions of spectral densities defined as follows.

Definition 1 (Wavelets)**.** *Let $\mathbb{N}_0 = \mathbb{N} \cup 0$. Let $\{\psi_{j,k}\}_{j\in\mathbb{N}_0,k\in\mathbb{Z}}$ be an orthonormal periodic wavelet basis for functions $f \in L^2(\Pi)$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, where $\psi \in L^2(\Pi)$ is a (suitably rescaled) Daubechies' compactly supported mother wavelet. The wavelet expansion of a spectral density can be written*

$$f(\omega) = \sum_{k\in\mathbb{Z}} v_k \phi_{0,k}(\omega) + \sum_{j\in\mathbb{N}_0} \sum_{k\in\mathbb{Z}} v_{j,k}\psi_{j,k}(\omega), \tag{4.2}$$

for $\omega \in \Pi$, where $\phi(\omega)$ is the scaling function associated with the wavelet basis. Define $< \cdot, \cdot >$ to be the usual inner product on $L^2(\Pi)$ given by $< f, g >= \int_{-\pi}^{\pi} f(\omega)g(\omega)d\omega$.

Wavelets provide a decomposition of a function, $f(\omega)$ which is localized in time and frequency. The wavelet coefficients sets for a wide variety of functions (technically in some Besov class) are sparse. This means that the essential information present in a spectrum can be represented in very few wavelet coefficients, generally fewer than presented in other representations such as Fourier or orthogonal polynomials. Our testing procedures exploit this sparsity, especially for the general wavelet test presented in Section 4.4. Wavelet transforms have further advantages in that their implementations are extremely fast and efficient and the coefficients provide useful information about location and scale of non-constancies. For further information on wavelets see Daubechies (1992) or Mallat (1998), Vidakovic (1999) or Nason (2008).

4.2.2 Assessing Spectral Constancy

Our approach investigates the constancy of the spectral density function $f(\omega)$ by examining its wavelet coefficients given by:

$$v_{j,k} = \sigma_X^{-2} \int_{-\pi}^{\pi} f(\omega)\psi_{j,k}(\omega)d\omega = \sigma_X^{-2} < f, \psi_{j,k} > \text{ for } j \in \mathbb{N}_0, k \in \mathbb{Z}. \tag{4.3}$$

For example, $\psi_{j,k}(x)$ might be the usual Haar wavelet system defined by $\psi^H_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ where

$$\psi(x) = \begin{cases} 1 & \text{if } x \in [0, 1/2), \\ -1 & \text{if } x \in [1/2, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

The white noise null hypothesis can be reformulated in terms of the wavelet coefficients: if $f(\omega)$ is constant then every $v_{j,k} = 0$ and the null hypothesis is equivalent to $H_0 : v_{j,k} = 0$ for all $j \in \mathbb{N}_0, k \in \mathbb{Z}$. This is because a defining property of wavelets is that the integral of every wavelet is zero.

In practice, we do not know the spectrum and hence estimate it using the periodogram. For the white noise test we then apply the discrete Haar wavelet transform to the nor-

malised periodogram to form the following estimates of $v_{j,k}$:

$$\hat{v}_{j,k} = 2^{\frac{J-j}{2}} \hat{\sigma}_X^{-2} \left(\sum_{r=0}^{2^{J-j-1}-1} I_{2^{J-j}(k+1)-r} - \sum_{q=2^{J-j-1}}^{2^{J-j}-1} I_{2^{J-j}(k+1)-q} \right), \tag{4.5}$$

for $(j,k) \in \mathcal{I}_T = \{(j,k) : J = \log_2(T),\, j = 0, \ldots, J-1,\, k = 0, \ldots, 2^j - 1\}$. Here we assume $T = 2^J$ for some integer J, but the test can be extended to data sets by using the Haar wavelet transform for arbitrary-n. A version of (4.5) exists for more general Daubechies' wavelets and a fast transform, called the discrete wavelet transform, exists due to Mallat (1989).

To test constancy of f we perform multiple hypothesis tests for $H_0 : v_{j,k} = 0$ for all $(j,k) \in \mathcal{I}_T$ against $H_A : \exists (j,k)$ such that $v_{j,k} \neq 0$ using the $\hat{v}_{j,k}$ as test statistic. Sections 4.3 and 4.4 explain how to compute the distribution of $\hat{v}_{j,k}$ to a high degree of accuracy for both the Haar and general wavelet situation. From this we can obtain a 'per-test' p-value with respect to each j,k pair. We use multiple hypothesis test size adjustment techniques such as Bonferroni correction or the false discovery rate method of Benjamini and Hochberg (1995a) to obtain tests of size α from the multiple tests.

The idea of using Haar wavelet coefficients for testing constancy appeared in von Sachs and Neumann (2000) in the context of stationarity testing. Here we test for constancy over frequency, rather than over time, and use more general wavelets to improve the statistical power of our tests.

4.3 A Haar wavelet test

Under the null hypothesis of independent and identically distributed data it is well-known that the periodogram ordinates, I_p, are distributed asymptotically as independent exponential random variables with mean σ_X^2, see Brockwell and Davis (1991) page 344. If the X_t are Gaussian then the periodogram ordinates are also uncorrelated.

We will model the periodogram ordinates as independent exponential random vari-

ables with mean σ_X^2.

4.3.1 All coefficient Haar Test

For now suppose that $\sigma_X^2 = 1$ is known. Then under H_0 the distribution of $\hat{v}_{j,k}$ from (4.5) is given by the following proposition.

Proposition 2. *The distribution of $\hat{v}_{j,k}$ for $(j,k) \in \mathcal{I}_T$ is given by*

$$g_m(x) = \frac{\sqrt{2m}\exp(-\sqrt{2m}|x|)}{2^{2m-1}(m-1)!}\sum_{j=1}^{m}\frac{(m+j-2)!}{(m-j)!(j-1)!}\left(2\sqrt{2m}|x|\right)^{m-j}, \tag{4.6}$$

where $m = 2^{J-j-1}$ for $j = 0, \ldots, J-1$ and $\sigma^2 = 1$.

Proof: see appendix. We call $g_m(x)$ Macdonald's distribution although $g_1(x)$ is the scaled Laplace distribution with variance one. In the Kotz et al. (2001) chapter 4, a general form of this distribution is called the Bessel distribution.

In practice σ_X^2 is unknown and so we estimate it using the standard sum-of-squares sample variance formula. Then we operate our test on the normalized periodogram $\hat{\sigma}_X^{-2} I_p$ which is distributed, asymptotically, as an exponentially distributed random variable with rate one.

We will abbreviate the all coefficient Haar wavelet test by *HWWN*.

Test HWWN procedure. To carry out the test we compute:

1. the normalized periodogram

2. then compute its Haar wavelet transform as specified by (4.5).

3. The Haar coefficients have distribution specified by (4.6) and the p-value for testing $H_0 : v_{j,k} = 0$ versus $H_A : v_{j,k} \neq 0$ can be obtained by using the inverse cumulative distribution of $g_m(x)$ from (4.6).

An important feature of the *HWWN* test is that no tuning parameters are required.

Figure 4.1: Top left: probability density estimate (solid) of Haar wavelet coefficients $\hat{v}_{11,k}$ with $g_2(x)$ superimposed (dotted). Top right: equivalent but for cumulative distribution. Bottom left: empirical histogram of p-values.

Figure 4.1 shows probability and cumulative density estimates of the wavelet coefficients with their theoretical $g_2(x)$ versions superimposed and associated empirical p-values obtained by the test procedure operating on 2^{14} independent Gaussian random variables. The empirical p-values appear to be close to uniform random variables which is what one would expect for the test statistic under the null hypothesis.

The result in Proposition 2 is approximate for the null distribution of $\hat{v}_{j,k}$. For a finite sample the periodogram ordinates are only approximately independently exponentially distributed. However, even for small T the approximation appears to be good. Furthermore, any small correlations in the periodogram ordinates are likely to be reduced further by the well-known 'decorrelation property' of wavelet transforms, see Johnstone and Silverman (1997), for example. A second factor is that the finite wavelet coefficients in (4.5) are normalized by the sample variance $\hat{\sigma}_X^2$ which causes normalized periodogram ordinates to obey a F-distribution, see Koen (1990). However, in practice, the difference between this F-distribution and the exponential is very small indeed, even for moderate sample sizes.

4.3.2 Single coefficient Haar test

Any non-zero Haar coefficient in the test described above indicates departures from white noise. We develop a second test based on the single coefficient $v_{0,0}$, which we relabel $d_{0,0}$ to indicate that it forms a different test. The coefficient $d_{0,0}$ is the coarsest scale wavelet coefficient and can be written in the frequency domain as:

$$d_{0,0} = 2^{-1/2}\sigma_X^{-2}\left(\int_0^{\pi/2} f(\omega)\,d\omega - \int_{\pi/2}^{\pi} f(\omega)\,d\omega\right). \tag{4.7}$$

Due to the Fourier relationship between autocovariance and spectrum (and some algebra) the coefficient $d_{0,0}$ can be rewritten as

$$d_{0,0} = (4/\sqrt{2}\pi)\sigma_X^{-2}\sum_{m=0}^{\infty}\gamma(2m+1)/(2m+1) \tag{4.8}$$

$$\approx 0.90\rho(1) - 0.30\rho(3) + 0.18\rho(5) - 0.13\rho(7) + \cdots, \tag{4.9}$$

where $\rho(\tau)$ is the usual autocorrelation, see Appendix A.4.

A test statistic can obtained from $d_{0,0}$ by substituting the sample autocorrelation, r_τ for $\rho(\tau)$ and comparing $|\hat{d}_{0,0}|$ to a critical value obtained either from the Macdonald distribution from Proposition 2 or using a Gaussian approximation derived in the next section. Although this test is based on only one wavelet coefficient it is surprisingly effective in practice as it looks for gross scale departures of spectral constancy. The single coefficient test is also particularly easy to compute from the sample autocorrelation which, in many cases, would have already been computed and plotted.

4.4 A general wavelet test

A key property of wavelets is that wavelet expansions of a wide variety of functions are sparse, see Wasserman (2005, p. 208) or Nason (2008), for example. This is illustrated by the bottom-right plot in Figure 4.1 which displays the Haar wavelet coefficients of the AR(1) spectrum with $\alpha_1 = 0.9$. Most of the coefficients in the plot are small or zero and only four or five large ones are required to represent the spectrum. One reason for the sparsity of representation of wavelets can be explained using the vanishing moments property of wavelets. A wavelet $\psi(x)$ is said to possess m vanishing moments if $\int x^\ell \psi(x)\,dx = 0$ for $\ell = 0, \ldots, m-1$. This property means that coefficients of wavelets, derived from $\psi(x)$, that overlap the spectrum, where it is locally like a polynomial of degree less than m, will be small and in many cases zero. The consequence is that smooth spectral densities, such as those belonging to ARMA(p, q) processes, will be sparsely represented in wavelet bases and, in principle, the representation will be sparser when using wavelets

with a higher number of vanishing moments.

Hence, we extend the test from Section 4.3 by using Daubechies' compactly supported wavelets with ten vanishing moments in (4.3) instead of Haar wavelets. The discrete wavelet coefficients, associated with these smoother wavelets, can be written as $\hat{v}_{j,k} = \hat{\sigma}_X^{-2} \sum_i g_i^{(j,k)} I_i$ where I_i are the usual periodogram ordinates and the g_i are the weights of the discrete wavelet transform. Formula (4.5) can be put into this form with $g_i = \pm 2^{j/2}$ at scale j and then this window of coefficients is moved across the I_i sequence to obtain the $\hat{v}_{j,k}$ for k across this scale. Although the formula for computing $\hat{v}_{j,k}$, for all the $T/2$ periodogram coefficients, seems to require $\mathcal{O}(T^2)$ operations the discrete wavelet transform introduced by Mallat (1989) performs the transform in a remarkable $\mathcal{O}(T)$ operations.

For a white noise test we need to understand the distribution of $\hat{v}_{j,k}$ under the null hypothesis of constancy of the spectrum. For more general wavelets no *simple* closed form, such as in Proposition 2, seems to exist because the wavelet weights, g_i, are not constant and change from scale to scale. Alternatively, as long as we stay away from the finer scales, the wavelet coefficients become asymptotically normally distributed because of the central limit theorem.

The asymptotic normality can easily be established for the Haar wavelet case by the following argument. Proposition 2 shows that the characteristic function of the Macdonald distribution is essentially Student's t-distribution which is well-known to tend asymptotically to the normal distribution as the number of degrees of freedom tends to infinity. Hence, asymptotically, Macdonald tends to a normal distribution as the Gaussian density is a fixed point of the Fourier transform.

Next, we establish asymptotic normality for the $\hat{v}_{j,k}$ for general wavelets, specify a

general wavelet test, and then a formula for power function of a test based on the $\hat{v}_{j,k}$ at medium to coarse scales is derived.

4.4.1 Specification of the General Wavelet Test

We will establish the asymptotic normality of the $\hat{v}_{j,k}$ under the mild assumptions stated by Neumann (1996) as follows.

Assumption 2 (Cumulant Rate of Decay)**.** *Let X_t satisfy*

$$\sup_{1\leq t<\infty}\left\{\sum_{t_2,\ldots,t_k=1}^{\infty}|\operatorname{cum}(X_{t_1},\ldots,X_{t_k})|\right\}\leq C_1^k(k!)^{1+\gamma} \tag{4.10}$$

for all $k=2,3,\ldots$ where $\gamma\geq 0$.

Neumann (1996) Remark 3.1 notes that if $\{X_t\}$ is α-mixing with coefficients $\alpha(s)\leq K\exp(-b|s|)$ and $\mathrm{E}|X_t|^k\leq C^k(k!)^\gamma$ for all k then the equivalent bound in (4.10) is $C^k(k!)^{3+\gamma}$ and that many useful distributions, such as exponential, gamma, inverse Gaussian and the F-distribution show that the condition is satisfied for $\gamma=0$ and for more heavy-tailed distributions for $\gamma>0$.

Assumption 3. *The spectrum $f\in L^2(\Pi)$ of $\{X_t\}$ satisfies*

$$\mathrm{TV}(f)\leq C_2,\quad ||f||_\infty\leq C_3, \tag{4.11}$$

for some constants $C_2,C_3>0$, where $\mathrm{TV}(f)$ is the total variation of f. Additionally, we assume that $f\in B^m_{p,q}$ a Besov space.

For more details on Besov spaces see Abramovich et al. (1998) who note that "the parameter m measures the number of derivatives of f, the existence of derivatives is required in an L^p sense, whereas the parameter q provides a further finer gradation". Besov spaces provide the realm for a wide variety of functions including those with spatial inhomogeneities such as point discontinuities or other singularities. However, Besov spaces

can encapsulate Sobolev spaces $H^m = B^m_{2,2}$ which provide a natural realm for smooth spectra for well-known processes such as ARMA.

Assumption 4 (Wavelet Assumptions)**.** *For any $r > m$ assume that (i) $\phi, \psi \in C^r$, the space of continuous functions with r continuous derivatives; (ii) $\int \phi(\omega)\, d\omega = 1$; (iii) $\int \omega^k \psi(\omega)\, d\omega =$ for $0 \leq k \leq r+1$, i.e. the wavelet has r vanishing moments.*

In practice, we set r to be high, e.g. $r = 10$ vanishing moments as this general results in better sparsity of representation in (4.2) and faster progression to normality.

The asymptotic normality of the coefficients is now established.

Proposition 3 (Coefficient Asymptotic Normality)**.** *Let $\hat{v}_{j,k}$ be the empirical wavelet coefficients of the normalized periodogram computed by* (4.5) *for Haar wavelets or $\sum_i g_i I_i$, for other Daubechies compactly supported wavelets. Then asymptotically, as $T \to \infty$, for $2^j = \mathcal{O}(T^{-1/2})$, we have $\hat{v}_{j,k} \sim N(0,1)$ under H_0 and $\hat{v}_{j,k} \sim N(v_{j,k}, \eta^2_{j,k})$ under H_A, where $\eta^2_{j,k} =< \psi^2_{j,k}, \{\pi \sigma_X^{-2} f\}^2 >$, the coefficients of the squared wavelet transform of the square of the normalized spectrum.*

The proof and more details can be found in Appendix A.2. The $2^j = \mathcal{O}(T^{-1/2})$ ensures that we only consider coefficients away from the fine scales and is the same assumption (A1) from von Sachs and Neumann (2000).

4.4.2 Power Function of the General Test

This section establishes the power function of the general test.

Proposition 4 (Test Power Function)**.** *Let the nominal size of the test be α, the Bonferroni corrected size be $\alpha_c = N_c^{-1}\alpha$ and the Bonferroni critical value $C_{\alpha_c} = \Phi^{-1}(1 - \alpha_c/2)$, where Φ is the standard normal cumulative distribution function. Then the (approximate)*

power function of the test is given by

$$\begin{aligned}\mathbb{P}\{\textit{Rej } H_0|f(\omega)\} =& \mathbb{P}(C_{\alpha_c} < \max_{(j,k)\in\mathcal{I}_T} |\hat{v}_{j,k}|) \\ =& 1 - \prod_{(j,k)\in\mathcal{I}_T} \left\{\Phi_{\eta_{j,k}}(C_{\alpha_c} - v_{j,k}) - \Phi_{\eta_{j,k}}(-C_{\alpha_c} - v_{j,k})\right\}, \qquad (4.12)\end{aligned}$$

where $v_{j,k}$ and $\eta_{j,k}$ are given by (4.3) *and Proposition 3 respectively.*

For the proof, please see appendix A.3. Figure 4.2 illustrates the utility and accuracy of our theoretical power function by comparing it to simulation results from four of the models used in our simulation study below. The solid lines correspond to theoretical power computed using (4.12) and the circles correspond to the simulation results and their agreement is extremely good. The top left-hand plot in Figure 4.2 corresponds to independent and identically distributed standard normal variates and so the power function here is the Type 1 error or the statistical size. The nominal value for all these tests is 5% and the top left plot shows good agreement with this.

4.5 Computational Details

4.5.1 Implementation

Our tests are implemented in the `R` (R Development Core Team, 2009) programming language in our package `hwwntest`, Savchev and Nason (2015). The Haar wavelet test is called `hwwn.test` and the general wavelet test using the asymptotic normal approximation is called `genwwn.test`. In both cases the normalized spectrum is computed using the fast Fourier transform and the regular `var` variance function. The normalized spectrum is then subjected to a wavelet transform from the `wavethresh` package. The wavelet coefficients are then compared to Macdonald's distribution (for Haar) or the standard normal distribution (for general wavelets) and then the set of coefficient p-values is adjusted by a Bonferroni correction (or alternatively, other methods such as false discov-

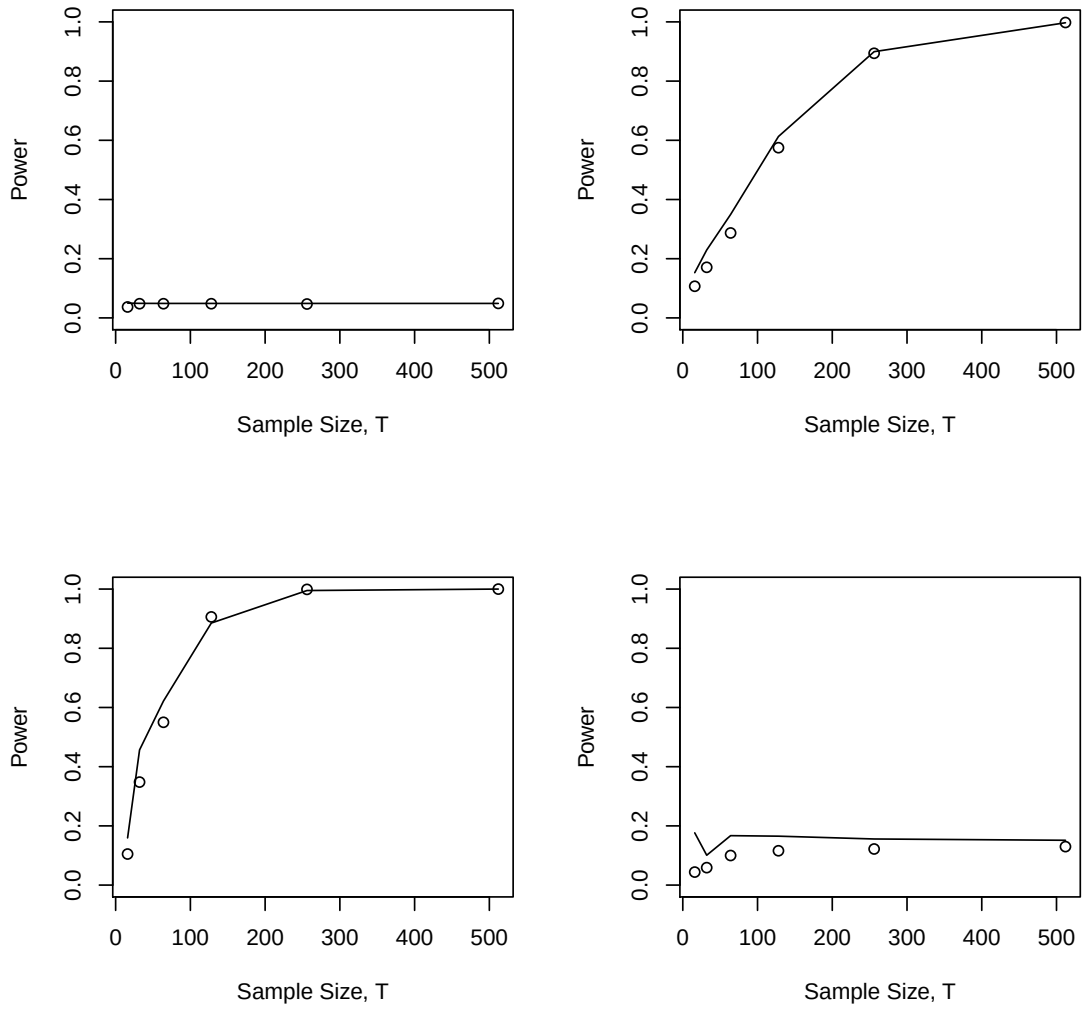

Figure 4.2: All plots: solid line is theoretical power calculated from (4.12) for the general wavelet test, circles are results from simulation study below. Top left: independent and identically distributed normal random variables; Top right: AR(1) with $\alpha = 0.3$, Table 4.2; Bottom left: MA(2) with $\beta_1 = 0, \beta_2 = 0.5$, Table 4.3, Bottom right: AR(12) with $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$, Table 4.3.

ery rate could be used).

To compute the theoretical power for a general spectrum we first compute the wavelet coefficients of the normalized spectrum. We then also need to calculate the variances $\eta_{j,k}^2 =< \psi_{j,k}^2, \{\pi\sigma_X^{-2} f\}^2 >$ which could be computed inefficiently by brute force methods. However, we choose to use the fast approximate methods due to Herrick et al. (2001) and Barber et al. (2002) where $\psi^2(x)$ is represented as a linear combination of father wavelets, $\phi_{\ell,m}$ at finer scales. Then all the rescaled squared wavelets, $\psi_{j,k}^2$ in the calculation of $\eta_{j,k}^2$ can be rapidly computed by using father wavelet coefficients at finer scales still.

Due to the use of fast wavelet algorithms operating on the periodogram the computational order of our test is $\mathcal{O}(T \log T)$. This order is shared by many of the other tests, e.g. the Bartlett and Ljung-Box tests share this speed. The fast speed can be important for applications that operate on long time series or situations where the test has to be applied many times. Also, the speed is an issue when comparing tests: e.g. if test A produces similar empirical power to test B, but operates slower by an order of magnitude, then test B is preferable, and, in some sense, test B is using the information more efficiently for the power assessment task.

4.5.2 Empirical distribution of the wavelet coefficients

The distribution of the Haar wavelets from (4.6) was derived under the Exponential assumption from Proposition 10.3.2 from Brockwell and Davis (1991) and the type of convergence is "in distribution". It is probably interesting to go in more numerical detail as to how the approximation is working and also with respect to the bias term. It is well known that the periodogram is an unbiased, but inconsistent estimator of the spectral density. If there is a bias it might be from the estimation of the wavelet coefficients which is based on the results in Neumann (1996), specifically Theorems 4.1 and 4.2 with the latter shown in appendix A.2 of the thesis which shows that the bias disappears in the asymptotic case.

Indeed Neumann (1996) Proposition 3.1(i) states that under the Assumptions 1, 2, 3 and 4 from Chapter 4 of the thesis, the expectation of the empirical wavelet coefficients is:

$$\mathrm{E}(\hat{v}_{j,k}) = v_{j,k} + \mathcal{O}(2^{j/2}T^{-1}log_2 T) \tag{4.13}$$

So, when using coarser scales this bias term from equation (4.13) goes to zero with a rate of $\mathcal{O}(T^{-1})$. Let us consider the finest scale for having T=128 observations, this means 64 for the periodogram, so the bias would be around 0.5, which is maybe a bit high with respect to the finest-scale wavelet coefficients, however not near any large value that would result in a type two false negative error.

With regard to the large number of observations shown on the graph on figure 4.1, we will now show it step by step with smaller ranges. Furthermore, we will do a more detailed simulation showing how the empirical distribution from data of the finest-scale Haar wavelet coefficients compare with the theoretical Macdonald distribution from equation (4.6) from p.73 in the thesis. If this bias is large enough, it should be shown from the empirical distribution.

Since we have T datapoints and $T/2$ for the periodogram's positive Fourier frequencies, then the number of finest scale wavelet coefficients would be $T/4$. Figure 4.3 shows with 10^3 realizations, that even for $T = 128$, which results in only 32 finest scale coefficients, there is no inherent bias present. Fig. 4.4 shows 10^3 realizations of Gaussian white noise using the `rnorm` function in `R` with the standard normal curve superimposed with red. The dataset sizes of $32, 64, 128, 256$ correspond to the number of wavelet coefficients from Fig. 4.3. We can notice that the convergence to the theoretical Gaussian curve seems to be the same or no worse than the one of the empirical wavelet coefficients to the Macdonald distribution. Furthermore, it is probably worth mentioning that the standard Macdonald/Bessel distribution that we are using converges to the standard Gaussian as $T \to \infty$ drives the degrees of freedom parameter $m = 2^{J-j-1}$ similarly to the way the t-distribution converges to the Normal distribution, Kotz et al. (2001) chapter 4. Fig.

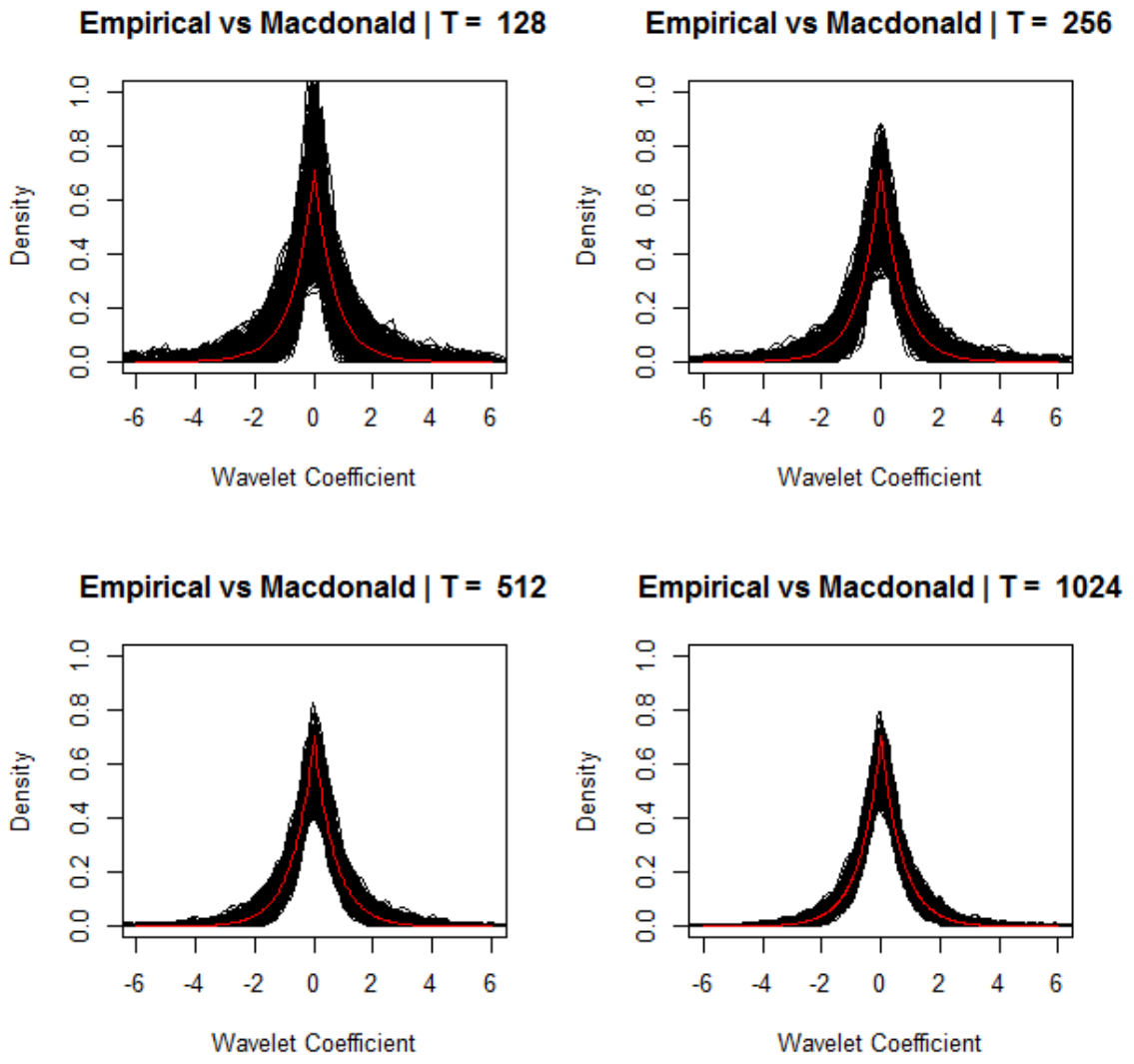

Figure 4.3: Top left to bottom right respectively: black — Empirical distribution of finest-scale Haar wavelet coefficients of 10^3 realizations from Gassian white noise with $T = 128, 256, 512, 1024$, red — theoretical Macdonald curve

.

4.5 shows the Macdonald distribution, starting from Laplace distribution ($m = 1$, finest scale $j = J - 1$). From the medium scale $j = J - 3$, the distribution is very close to the standard Gaussian. Moreover, both the t and the Macdonald/Bessel distribution belong to the class of generalised hyperbolic distribution, thus the Cauchy as well, which is the reason for the good empirical size with Cauchy white noise reported in section 4.6.1.

We have also run simulations with different values of σ and the results are no worse than the ones with $\sigma = 1$, since in the HWWN procedure we are using the normalised periodogram $I\hat{\sigma}^{-2}$. With respect to time-varying σ, our tests do not do well since our null hypothesis is not time-varying white noise under. With respect to the long-run variance as meant to a way of expressing σ^2 as a linear combination of autocovariances, it is true that we do not observe all covariances since we estimate the spectrum with the raw periodogram at the Fourier frequencies. Indeed, a kernel or tapered estimate of the periodogram might improve the statistical power under serial correlation alternatives and that would in itself be an interesting direction of new methodological research.

4.6 Univariate White Noise Simulation Study

We carried out an extensive simulation study to evaluate the empirical size and power of our new tests in comparison with the Box-Ljung, Bartlett and `whitenoise.test()` test from Lobato and Velasco (2004). The study results appear in Tables 4.1 to 4.4. The nominal size of all hypothesis tests was set to 5% and results established via 10^5 replications. A larger selection of results for a wider range of sample sizes and models can be found in Nason and Savchev (2014a).

4.6.1 Size Estimation for the three Wavelet Tests and Others

Table 4.1 presents empirical size results for data that are independent and identically distributed for both standard normal and Cauchy distributions for sample sizes of $T =$

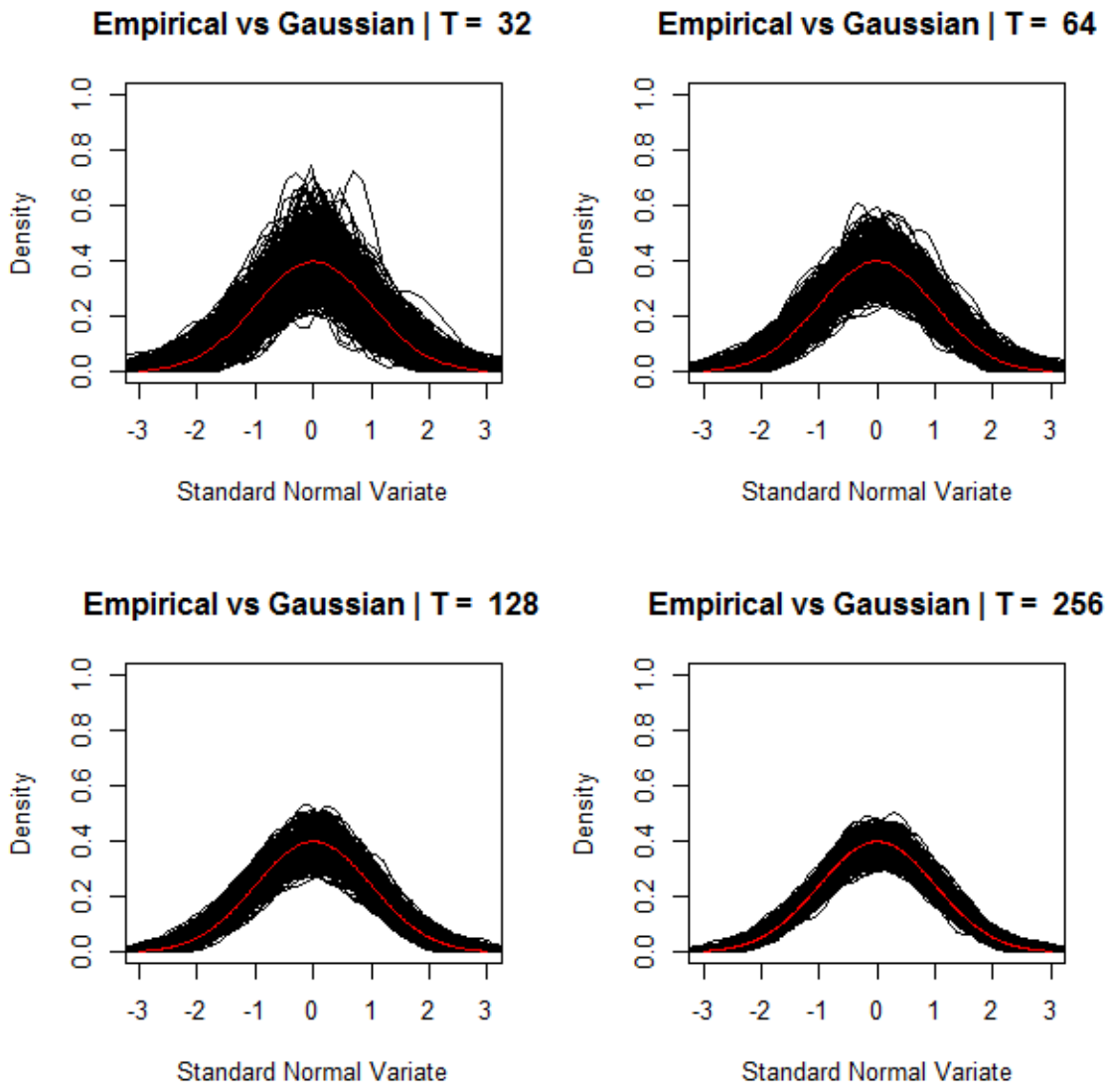

Figure 4.4: Top left to bottom right respectively: black — Empirical distribution of 10^3 realizations from Gassian white noise with $T = 32, 64, 128, 256$, red — standard Gaussian curve $N(0, 1)$

.

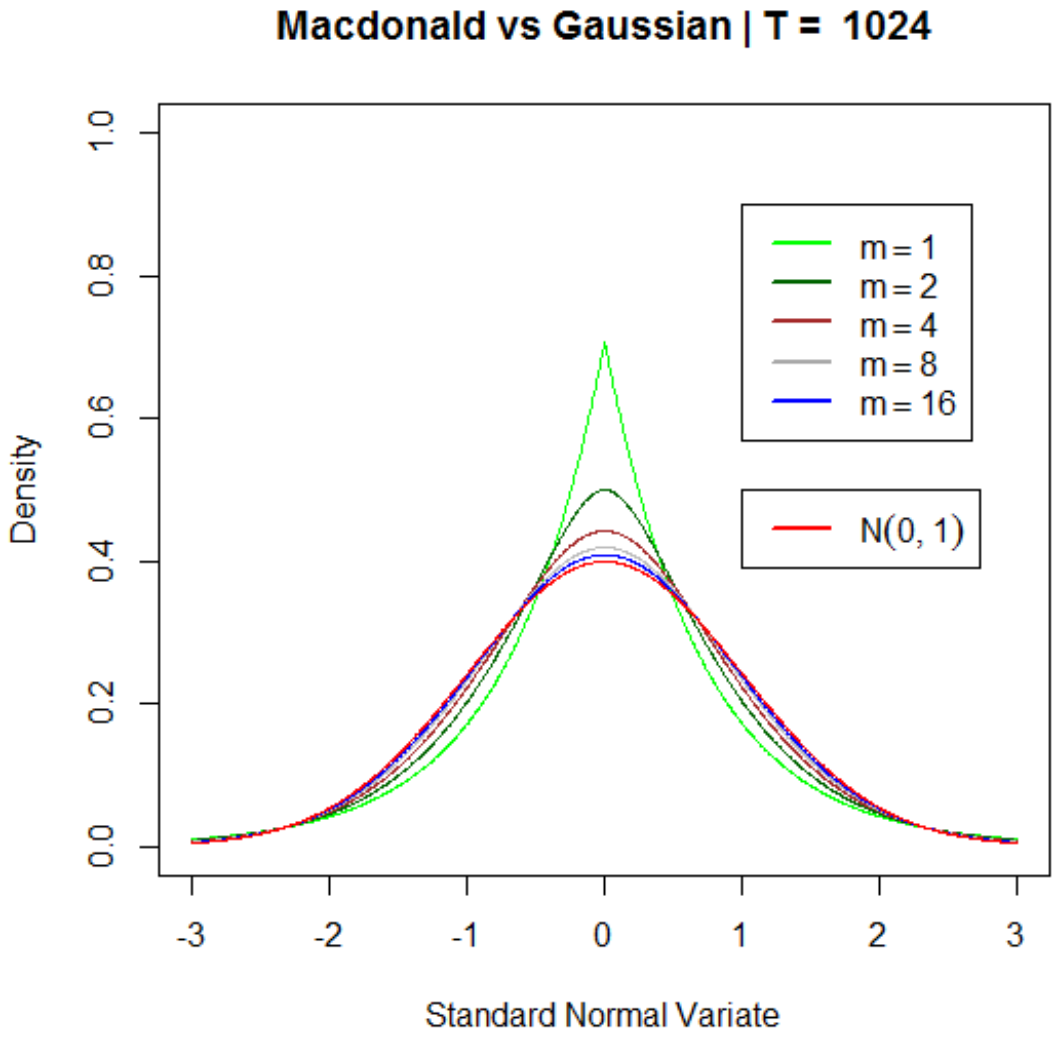

Figure 4.5: Shapes of the Macdonald distribution with varying the m parameter; red — standard Gaussian curve $N(0, 1)$

.

Model	T	hwwn	genwwn	Box		Bartlett	normwhn
				lag=1	lag=20		
$N(0,1)$	64	3.0	4.8	4.6	2.8	2.9	4.7
	256	3.8	4.7	4.9	4.3	3.7	4.7
	1024	4.1	4.9	4.9	4.9	4.1	4.9
Cauchy	64	0.4	2.0	2.7	0.3	1.5	30.7
	256	1.0	5.0	2.7	5.6	2.0	63.8
	1024	2.0	5.0	1.9	7.5	2.1	87.3

Table 4.1: Empirical size for the five white noise tests for various sample sizes, T. True model is independent and identically distributed variates. Approximate theoretical power from (4.12) computed to be 4.9% for all T for the General Wavelet Test genwwn.

$64, 256$ and 1024. In most cases the empirical size is close to, and never exceeds, the nominal size for all tests for the standard normal data. However, for Cauchy data, the empirical size for the normwhn test dramatically exceeds its nominal value and hence we would not recommend its use in scenarios where heavy-tailed noise might be present. There is also some evidence that the Ljung-Box test exceeds its nominal size for the Cauchy data. Tables 2 and 4 of Nason and Savchev (2014a) show further results for sample sizes $T = 16, 32, 128$ and 512 for Student's t-distribution variates on two and three degrees of freedom and also show the normwhn test performing poorly for these other heavy-tailed distributions. The periodogram based tests (hwwn, genwwn and Bartlett's test) are not affected presumably because of the fact that a periodogram's asymptotic independent and identically distributed exponential distribution happens under mild distributional conditions.

4.6.2 Power Estimation for the three Wavelet Tests and Others

Tables 4.2 to 4.4 show empirical power results for different alternative models with Gaussian innovations.

Table 4.2 compares three AR(1) models where there is not a lot of difference in the performance of the tests for $\alpha = \pm 0.9$ except that normwhn is less powerful for $\alpha = 0.9$. For $\alpha = 0.3$ the Ljung-Box test with lag=1 dominates, but this is not surprising since there is a large lag one coefficient. In practice, of course, the best lag is not known and,

Model	T	hwwn	genwwn	Box		Bartlett	normwhn
				lag=1	lag=20		
$\alpha = 0.3$	32	6.1	17.1	26.0	4.4	12.5	7.9
	64	16.2	28.7	58.1	18.0	41.6	12.9
	256	79.1	89.4	99.7	83.7	98.7	40.7
	1024	100.0	100.0	100.0	100.0	100.0	91.6
$\alpha = 0.9$	32	86.0	95.5	98.9	84.3	96.3	61.1
	64	100.0	100.0	100.0	99.9	100.0	84.0
$\alpha = -0.9$	32	96.0	98.0	99.9	95.2	99.7	96.9
	64	100.0	100.0	100.0	100.0	100.0	100.0

Table 4.2: Empirical power for the five white noise tests for various sample sizes, T. True model is AR(1) with parameter α with standard normal innovations.

e.g., for `lag=20` the `genwwn` and Bartlett test dominate the Ljung-Box test, with Bartlett being better for larger T.

Table 4.3 compares three ARMA models. The results for Box-Ljung vary depending on the supplied `lag`. The general wavelet test `genwwn` dominates for the MA(2) model, the Bartlett test dominates for the ARMA$(1, 2)$ model but the `genwwn` general wavelet test does well for larger T. For the AR(12) test the `normwhn` test performs well, but the Haar wavelet test `hwwn` is second best, and the Bartlett tests poorly.

Table 4.4 reruns some of the power simulations with heavy-tailed innovations. The general wavelet test, `genwwn`, works particularly well for Models a. and b., but not for c. However, the Haar wavelet test `hwwn` works well for model c. Here we discount `normwhn` as it has poor size characteristics for heavy tailed noise.

Overall conclusions from our study are (i) unless one was absolutely sure that their time series was not heavy-tailed, one should probably avoid the `normwhn` test, (ii) performance of the Box-Ljung test is highly variable and dependent on the number of lags considered, (iii) all of the tests perform well against some alternatives and not against others. Generally, the `genwwn` test performs creditably well.

Model	T	hwwn	genwwn	Box		Bartlett	normwhn
				lag=1	lag=20		
MA(2)	64	26.7	55.0	17.0	34.7	38.3	29.8
	256	93.6	99.9	17.8	99.8	98.8	77.0
	1024	100.0	100.0	17.9	100.0	100.0	100.0
ARMA(1, 2)	16	42.5	19.4	93.3	$*$	80.5	58.6
	32	93.1	98.0	100.0	79.1	99.7	87.6
	64	100.0	100.0	100.0	100.0	100.0	99.3
AR(12)	64	12.7	10.0	7.4	24.1	6.0	18.9
	256	51.9	12.2	8.9	98.6	14.6	75.9
	1024	98.9	13.2	9.4	100.0	36.1	100.0

Table 4.3: Empirical power for the five white noise tests for various sample sizes, T, all with standard normal innovations. MA(2) model has $\beta_1 = 0, \beta_2 = 0.5$, ARMA$(1, 2)$ model has $\alpha_1 = -0.4$, $\beta_1 = -0.8, \beta_2 = 0.4$, AR(12) model has $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$.

Model	hwwn	genwwn	Box		Bartlett	normwhn
			Box.lag=1	Box.lag=20		
a.	81.2	90.5	99.4	76.0	98.8	29.2
	79.1	*89.4*	*99.7*	*83.7*	*98.7*	*40.7*
b.	93.4	99.7	12.9	99.2	98.4	58.7
	93.6	*99.9*	*17.8*	*99.8*	*98.8*	*77.0*
c.	39.4	7.2	6.7	96.6	10.2	60.0
	51.9	*12.2*	*8.9*	*98.6*	*14.6*	*75.9*

Table 4.4: Empirical power for the three white noise tests for $T = 256$ for different models with Student's t distributed noise with two degrees of freedom in roman font. Results with Gaussian innovations, reproduced from tables above, are in italic font. Model: a.) AR(1) $\alpha = 0.3$ from Table 4.2; b.) MA(2), $(\beta_1 = 0, \beta_2 = 0.5)$ from Table 4.3; c.) AR(12), $\alpha_1 = \cdots = \alpha_{11} = 0$, $\alpha_{12} = -0.4$ from Table 4.3.

4.6.3 Simulation Comparisons with Contemporary Papers

We also compared some simulation results to existing literature. For example, with independent and identically distributed χ^2_1 random variables, with sample sizes of $T = 256$ and 1024 all our simulations for both wavelet tests, the Ljung-Box test, the Bartlett test and the whitenoise.test() were all less than 5% with the smallest being 2.91%, apart from the Ljung-Box test with lag $p = 10$ where the size was 5.1% and 5.06% for both sample sizes respectively. Hence, size control for all these tests is, in general, very good. Table 2 from Guay et al. (2013) reports results on equivalent models where their tests are shown to have empirical size values quoted of 5.61%, 5.83% for their *BP* test and 4.7% and 5.17% or their *Parz* for sample sizes of $T = 200$ and $T = 1000$, respectively. They also compare their tests to three others: the EL_n test by Escanciano and Lobato (2009), an adapted form of the Newey-West data-driven statistics used by Hong and Lee (2005) and the *CvM* test from Deo (2000). The empirical sizes for these tests are, for $T = 200$, 8.7%, 10.37% and 6.32% and, for $T = 1000$, 7.18%, 10.37% and 5.76% which are maybe a bit high.

Guay et al. (2013) perform further empirical power simulations with AR(1) and MA(1) models with challenging $o(n^{-1/2})$ parameters. Their results for MA(1) model, for the tests mentioned above, give a best power of 10.76% for $T = 200$ and 34.4% for $T = 1000$ whereas the $d_{0,0}$ single coefficient from Section 4.3.2 achieves a competitive power of 10.49% and 28.64% and, of course, the $d_{0,0}$ test is very simple, requiring no tuning parameters and directly uses the sample autocorrelation for computation. For the AR(1) model in their Table 4 the maximum power for $T = 200$ is 10.5% and for $T = 1000$ it is 35.1% and $d_{0,0}$ gives 10% and 29.64% respectively.

Guay et al. (2013) also use a MA(4) and AR(6) model mimicking "hidden periodicity" in their Table 4. Our general wavelet test achieves extremely high empirical power for these models achieving 41% and 92% respectively whereas the several tests in Guay et al. (2013) do not have power greater than 20% and 70%. Furthermore, our test has no tuning

parameters whereas tests in Guay et al. (2013) combine results from lags ranging from 1.9 to 35.2 — those lags are not integer due to Guay et al. (2013) averaging procedure for obtaining the final test result.

4.7 Real Data Examples

4.7.1 Wind Speed Example

Figure 4.6 shows a time series of $T = 128$ observations from a larger set of hourly wind speeds recorded at Aberporth, Wales during 2010. The top left plot shows the actual values of the series and the top right plot the first differences of the series which detrends the original. It is of interest to determine whether there is a non-white noise structure within the first differences to aid forecasting. Such short term forecasts are of value for companies running wind farms and power companies to enable them to aggregate power from a variety of sources efficiently. The autocorrelation function of the first differences is plotted in the bottom left-hand plot of Figure 4.6 and one can see that it is close to white noise, although there might be significant lags at $\tau = 1$ and $\tau = 10$. Other tests indicate that there is little evidence of heavy tails.

First, we consider the Ljung-Box test. The p-value for this test with the lag $p = 1$ is 0.004 which indicates extremely strong evidence to reject H_0 of white noise. On the other hand, the Ljung-Box test for $p = 20$ gives a p-value of 0.08 which indicates that there is no evidence to reject H_0. Other values of the Ljung-Box lag give a variety of other p-values which makes it impossible to know whether to reject H_0 or not.

The p-values for the other tests are 10^{-15} for the `whitenoise.test` from Lobato and Velasco (2004), 0.026 for the Bartlett test, 1.00 for the Haar wavelet test, 0.011 for the single coefficient $d_{0,0}$ test and 0.044 for the general wavelet test. The cumulative normalized periodogram used to compute the Bartlett test statistics is the bottom right plot of Figure 4.6. Our practical conclusion is that there is strong evidence that the series is not white noise and we reject the null hypothesis.

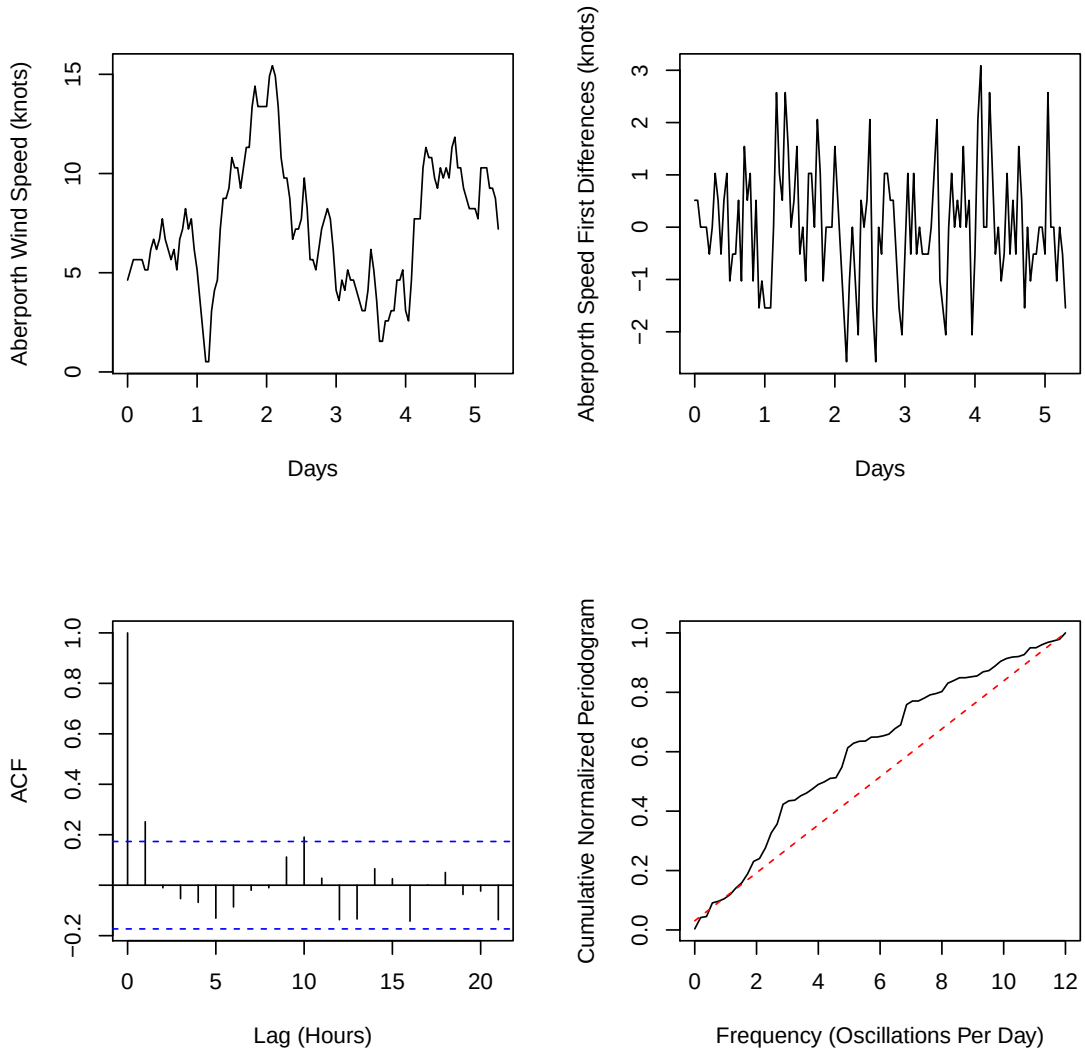

Figure 4.6: Top left: Aberporth wind speed time series. Top right: first differences of Aberporth wind speed time series. Bottom left: Autocorrelation function of Aberporth wind speed first differences. Bottom right: Cumulative normalized periodogram (solid black) and ideal white noise line (red dotted) for Bartlett white noise test.

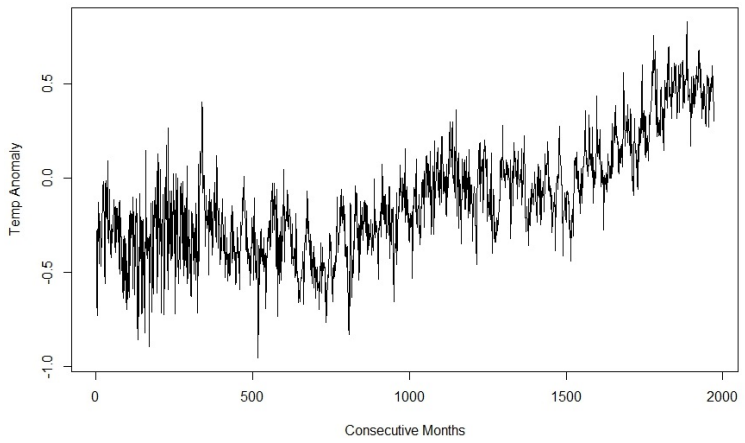

Figure 4.7: Hadcrut4: Global ensemble medians of temperature anomalies from 01/1850 till 04/2014

4.7.2 The HADCRUT4 Global Dataset

Description of the Data

Here we examine an Instrumental Temperature Record, please see Jones et al. (1999). We will be using data up to April 2014 which is called *HadCRUT4*. These data represent digitised measurements which include the climate error models for uncertainty on a scale that represents temperature anomalies relative to 1961–1990 with data from 1850 till today, Jones et al. (1999). We will use HadCRUT4 time series: global ensemble medians and uncertainties for the global level which average southern and northern hemisphere estimates on a monthly level. The data can be downloaded from the link: http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html.

Our Analysis

For our analysis, we will use the first 1025 monthly observations from the data which corresponds with January 1850 until April 1935. They are plotted in Fig. 4.8.

Fig. 4.9 shows that the partial and regular autocorrelation of the total series was not decaying. This pattern is largely preserved in Fig. 4.10 — for our chosen subset of monthly data and we will work with the lag one differenced data. Fig. 4.11 shows the

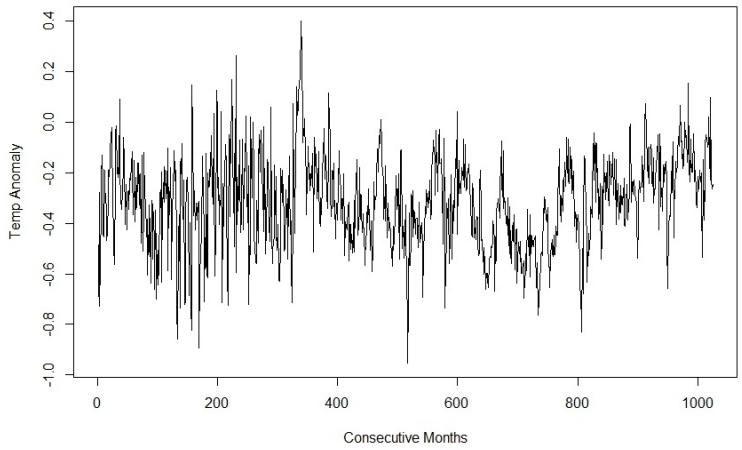

Figure 4.8: Hadcrut4: Global ensemble medians of temperature anomalies from $01/1850$ till $04/1935$ (first 1025 observations of the data from Fig. 4.7

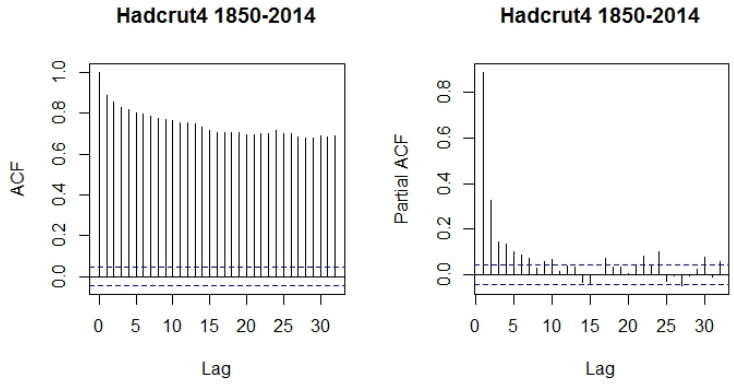

Figure 4.9: ACF and PACF of the temperature anomalies monthly raw data from $01/1850$ till $04/2014$.

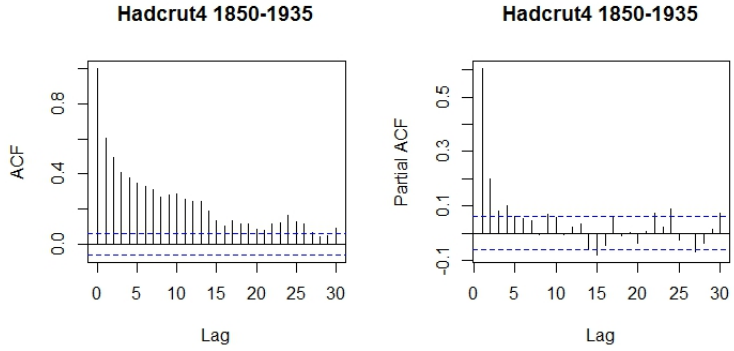

Figure 4.10: ACF and PACF of the temperature anomalies monthly raw data from $01/1850$ till $04/1935$.

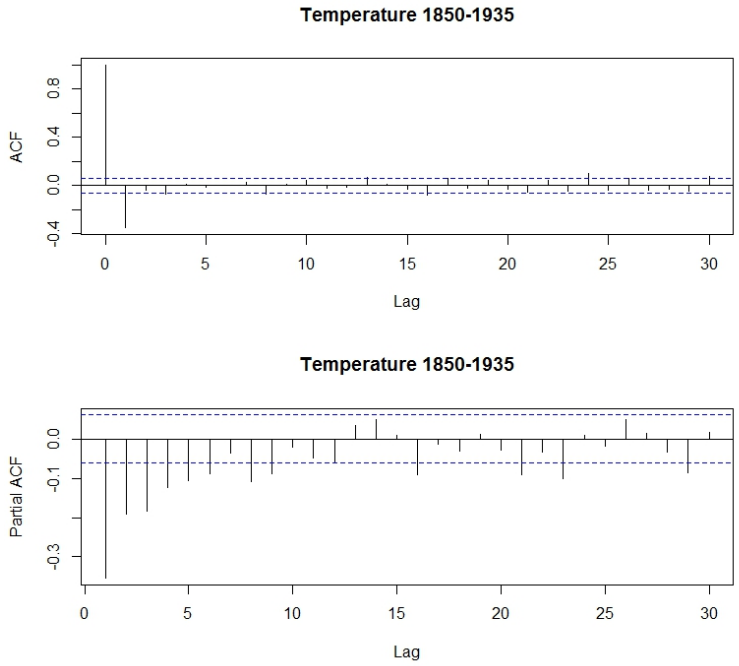

Figure 4.11: ACF and PACF of first order diffences of the temperature data.

auto and partial correlations of our 1025 monthly subset after differencing one time. They suggest that a mixed ARMA model would be suitable. After consideration of several models, based on AIC and 100 one-step-ahead forecasts, the best model we chose was ARMA $(1, 1)$ for the differenced data. Hence we model the anomaly data from $01/1850$ till $04/1935$ by an ARIMA$(1, 1, 1)$ model .

Now we turn to residual analysis and application of several white noise tests. The idea is that we hope to have captured most of the variance with our model. However, as explained, this dataset describes a complex natural phenomenon, so there might be hidden higher lag temperature periodicites that have been missed by the model. We will apply our wavelet tests *(Haar, General and d00)*, the tests of Bartlett and Lobato, and Ljung-Box. Table 4.5 shows that the Haar wavelet test and the Lobato test reject the null hypothesis. This is pointing in the direction that there is a hidden periodicity. We also performed the Ljung-Box test for different lags to see if it would agrees with our findings. From the results in table 4.6, we can infer that the hidden periodicity is between lag 12 and lag 18

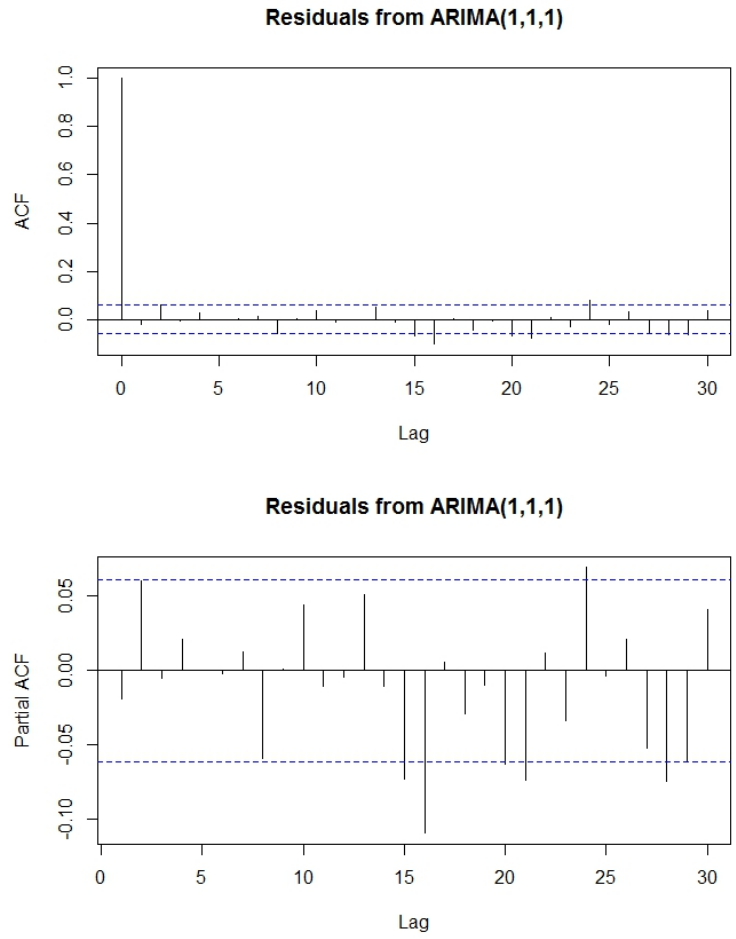

Figure 4.12: Auto and Partial (auto) correlations of the residual series from ARIMA(1,1,1) model on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset

as the L-B test rejects null hypothesis in this region. This confirms our assertions from the simulation study in the previous section that the wavelet periodogram tests are suitable for such a task. Indeed, if we look more closely at the autocorrelations of the residuals from our model on Fig. 4.12, we can notice that both ACF and PACF are bordering and crossing the confidence interval at lags 15 and 16 respectively. This finding might be a random effect, though in climate data, there often are certain hidden periodicities or complicated cycles. To sum up, our Haar wavelet test agrees with the tests of Lobato, and Ljung-Box, when we specify the lag correctly. The test of Lobato is implemented in the `normwhn` `R` package.

Test	hwwn	genwwn	d00	Bartlett	normwhn
p-value	0.0028	0.7298	0.7982	0.2952	0.0020

Table 4.5: Results from applying white noise tests to the residuals of ARIMA(1,1,1) on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset

L-B Test	lag 6	lag 12	lag 18	lag 24
p-value	0.1823	0.3441	0.0087	0.0004

Table 4.6: Results from applying the Ljung-Box test with different lags to the residuals of ARIMA(1,1,1) on the 1025 monthly obsevations — from 1850 till 1935 — of the HADCRUT4 Global Dataset

4.7.3 S&P 500 Annual Log Returns

Description of the Data

The dataset we are going to analyze in this section are the annual returns from Standard and Poor's 500 (S&P), which is a major US stock market index of the top 500 companies. We will use data from 1871 till 1998 which totals 128 observations. The data from 1871 till 1957 (when the index was officially introduced) are calculated from the top 500 companies at the respective year. The source is: http://data.okfn.org/data/core/s-and-p-500

Those data are also analyzed in Lobato (2001).

Analysis and Test Results

We decided to use the natural logarithm of the ratio of the current year to the previous one or what is known as the *log-returns* or the first differences of the logarithms of the raw data. This is a typical transformation for such types of data since financial returns are believed to be log-normally distributed, Baxter and Rennie (1996) page 6. The data we use are shown on Fig. 4.13, where the big downward spike in the middle occurss at the time of the Great Depression. Another basic idea in financial time series is that the log-returns are usually not serially correlated, Samuelson (1965). On Fig. 4.14 we inspect

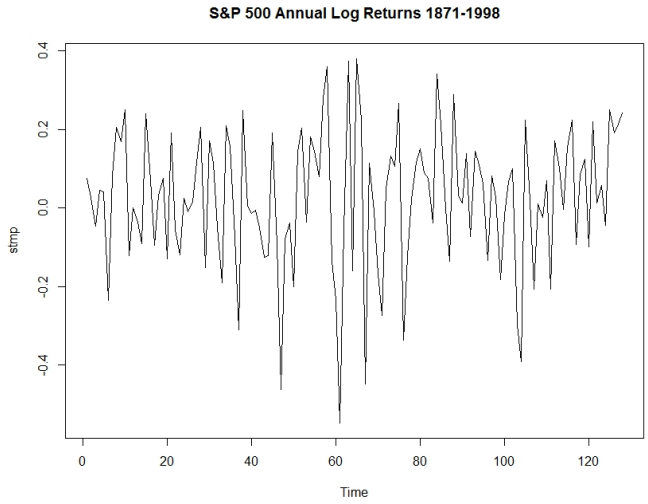

Figure 4.13: Annual Log-Returns from S&P 500 from 1871 till 1998

Test	hwwn	genwwn	d00	Bartlett	normwhn
p-value	0.2882	0.1054	0.6959	0.2836	0.6748

Table 4.7: Results from applying white noise tests to the log-returns from S&P 500 for 1871-1998

the sample (and partial) autocorrelation function. We notice that most of the correlations are within the confidence intervals, though lag two is on the borderline. However, this is pretty common for white noise data. Overall, the correlogram looks pretty close to white noise.

Now, looking at tables 4.7 and 4.8, we can conclude that all mentioned tests do not reject the null hypothesis of white noise. The conclusion in Lobato (2001) is the same, using their test and another modified Box-type test.

L-B Test	lag 1	lag 2	lag 3	lag 10
p-value	0.6556	0.1173	0.1117	0.3415

Table 4.8: Results from applying the Ljung-Box test with different lags to the log-returns from S&P 500 for 1871-1998

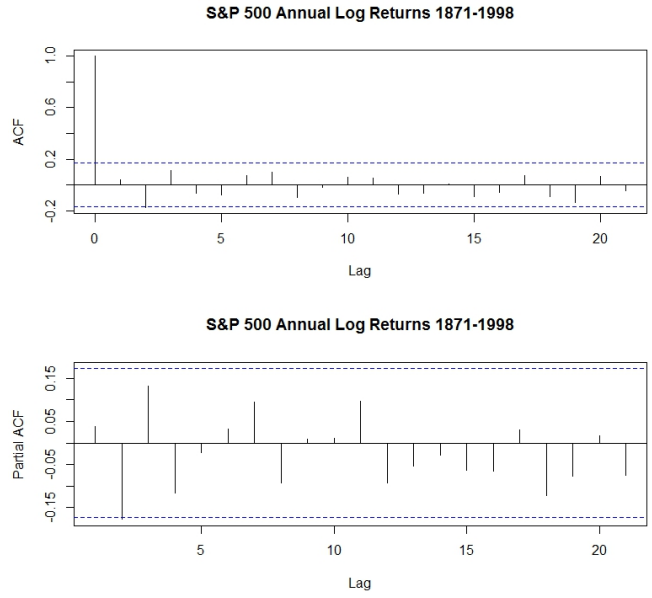

Figure 4.14: ACF and PACF of the lLog-Returns from S&P 500 from 1871 till 1998

Chapter 5

Local Alternatives and Nonlinear Models

5.1 Introduction

The partial differences of average of Fourier periodograms over consecutive Fourier frequencies in white noise testing have a different interpretation than stationarity testing as in Nason (2013). The essential problem underlying the wavelet white noise tests in the thesis is testing for constancy of the spectrum over Fourier frequencies or more precisely over 'scales of frequency'. By scales, it is meant that for the ARMA alternative we might have low or high-frequency processes as well as mid-low and mid-high ones and also mid-frequency where the peak of the process in around $\pi/2$. Moreover, for higher orders such as AR/MA(4) with negative parameter, the spectrum would have two peaks — one around $\pi/3$ and the other at $2\pi/3$. This means that depending on the alternative hypothesis, some fine-tuning of our tests might produce better results.

For instance, considering the ARMA alternative hypothesis, we might have processes with high or low frequencies peaks, for which d_{00} would be one of the best. However, there are also cases where the peak of the spectrum is in the mid-frequencies range around $\pi/2$. It is explained in appendix A4 that similar to d_{00}, we can have d_{10} and d_{11}, then

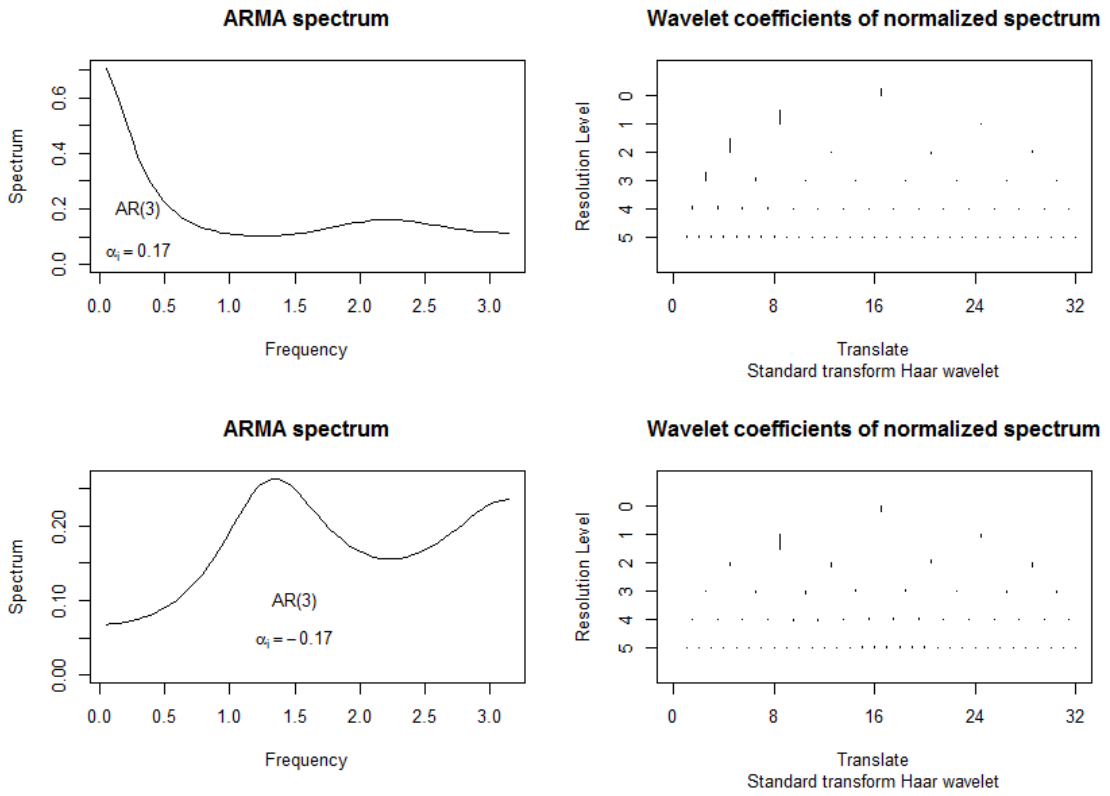

Figure 5.1: Top left: Spectrum of AR(3) process with parameter $\alpha_i = 0.1767$ for $i = 1, 2, 3$ corresponding to local alternatives scenario $\alpha = 2/\sqrt{(T)}, \quad T = 128$. Top right: the Haar wavelet coefficients of the normalised spectrum with $T = 128$. Bottom left and right: spectrum of AR(3) with negative parameters and its Haar wavelet coefficients for $T = 128$

.

$d_{20}, d_{21}, d_{22}, d_{23}$ etc. One such example is illustrated on Fig. 5.1. The scenario plotted in the top of Fig. 5.1 corresponds to a *local alternatives* scenario of order $\mathcal{O}(T^{-1/2})$, more precisely an AR(3) process with parameters $\alpha_i = 2/\sqrt(T)$ for $i = 1, 2, 3$ and $T = 128$. It is evident that the largest coefficients occur in scales $j = 2$ and $j = 3$. The approximate theoretical power for the d_{00} test is 46% with the coefficient $d_{00} = 1.84$ and a critical value of quantile $C_{\alpha_c} = 1.959$. Including $d_{10} = 3.48$ and $d_{11} = 0.24$ the power increases to 84%, despite using three coefficients and $C_{\alpha_c} = 2.39$. Furthermore, we can notice on Fig. 5.1 that the d_{20} is also large, it is 3.31. If we include all the 4 coefficients from scale $j = 2$ in the test, despite that the rest are smaller in magnitude than 0.27, the approximate theoretical power goes to 91% and the quantile $C_{\alpha_c} = 2.69$ since we are testing 7 coefficients altogether. In contrast, bottom of Fig. 5.1 shows the same local alternatives scenario with negative value of the AR parameter α which results in spectral peak in the mid frequencies around $\pi/2$. This time if we used only $d_{00} = -0.72$, the approximate theoretical power would be only 12% and when adding $d_{10} = -2.22$ and $d_{11} = 0.39$, the power is 46%. In comparison, in the first local alternatives scenario ($\alpha > 0$) of Fig. 5.1, results in higher power using scales $j = 2, 1, 0$, whereas in the case $\alpha < 0$ using scales only $j = 1, 0$ is better.

In summary, the parable is that there are probably more scenarios that could be easily caught by only the most global basis with the coefficient d_{00}, however there are also more subtle cases for which fine tuning is required such as considering the last three scales $j = 2, 1, 0$, leading to d_{11} and d_{22} tests. Moreover, it is probably surprising that in an easy case for AR(1), $\alpha = -0.9$ plotted on the bottom of Fig. 5.2 — a high frequency process — the largest Haar wavelet coefficient occurs in scale 3 rather than 0, which is well in-between global and local basis. Of course, in this case the $d_{00} = -8.64$ coefficient is also large, but $d_{37} = -14.1$. In the more moderate high frequency case on the top of Fig. 5.2 AR(1) with $\alpha = -0.6$. we have that the largest coefficient in magnitude is $d_{00} = -5.71$.

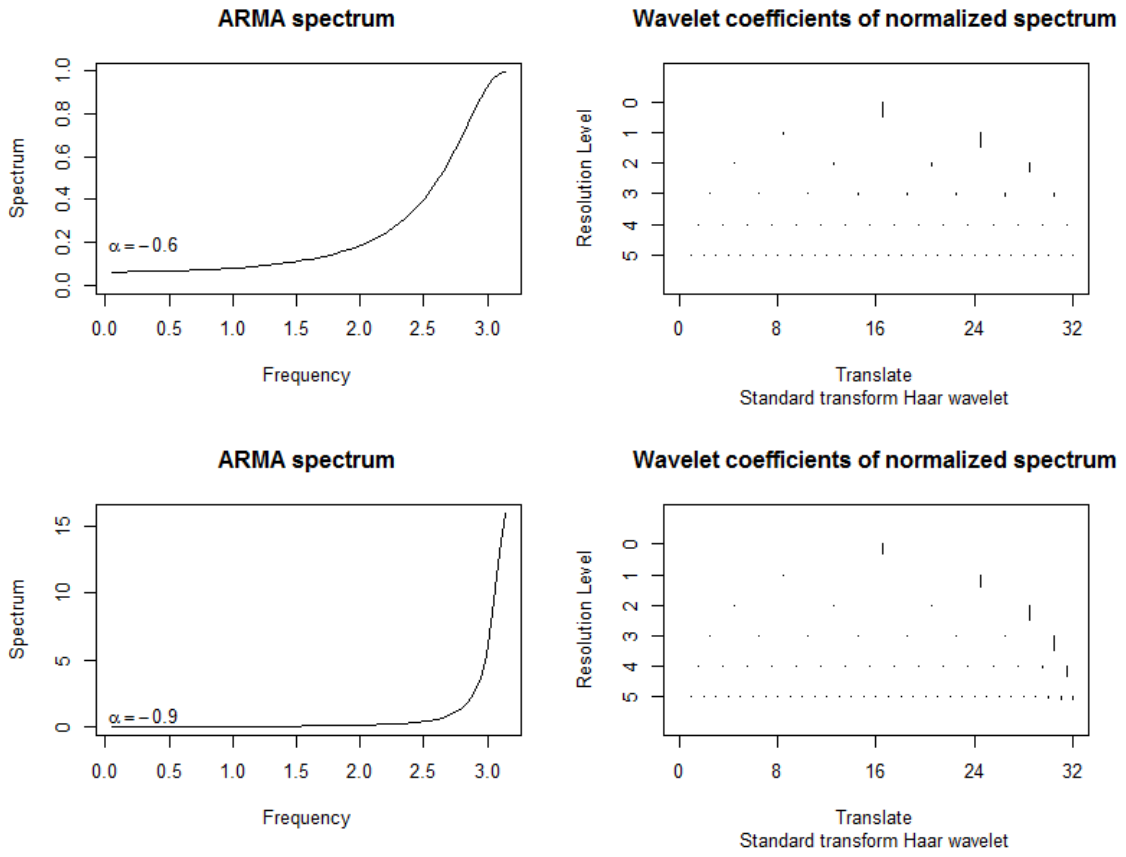

Figure 5.2: Top left: Spectrum of AR(1) process with parameter $\alpha_i = -0.6$. Top right: the Haar wavelet coefficients of the normalised spectrum with $T = 128$. Bottom left and right: spectrum of AR(1) with parameter $\alpha_i = -0.9$ and its Haar wavelet coefficients for $T = 128$

.

Last but not least, one of the practical contributions of this thesis is also a toolkit for implementing the tests and approximate theoretical calculations in our R package `hwwntest`, specifically the `genwwn.thpower` function which allows similar approximate calculations for different wavelets.

5.2 Local Alternatives

In this section we will look into practical application of the approximate theoretical power formula from Approximation 4, equation 4.12. In Engle (1984), a sequence of local alternatives is formed as follows:

$$H_a^T : \theta_1^T = \theta_1^0 + \delta/T^{1/2} \tag{5.1}$$

In our case $\delta \in \mathbb{Z}$. This type of local alternative (5.1) is useful to determine whether or not we have *invariant* test depending on the sign of δ. Ideally, a researcher would like to use invariant test, however not always possible.

For investigating our wavelet tests we suggest a scenario similar to the approach of Guay et al. (2013), however without the restrictions to large lags. In Guay et al. (2013), it is reported that the number of lags p which to consider for their test to detect such a local alternative might rise to $T^{1/2}$. Their condition is to have "sufficiently many such coefficients" In our case we suggest the following sequence of local alternatives:

5.2.1 AR/MA(p) local alternatives scenario

Let α_p and β_q for $p, q = 1, \ldots, \log_2 T$ be the parameters of an AR/MA process respectively. Additional restriction is that $\sum_{p,q=1}^{\log_2 T} \alpha_p + \beta_q < 1$, so in practice $\log_2 T$ may or may not be reached e.g. for $T = 128$ we can have at most 5 non-zero parameters of order $T^{1/2}$. Additionally, if α and β are with opposite signs and the same magnitude for the ARMA case, this would result in their cancelling out and producing flat spectrum. We

will consider the following scenarios.

1. Scenario 1 (AR):

   $H_0 : \alpha_p = 0$ and $H_a = \alpha_p = 2/\sqrt(T)$

   Scenario 1a: $H_a = -\alpha_p$

2. Scenario 2 (MA):

   $H_0 : \beta_q = 0$ and $H_a = \beta_q = 2/\sqrt(T)$

   Scenario 2a: $H_a = -\beta_q$

3. Scenario 3 (ARMA):

   $H_0 : \alpha_p = 0, \beta_p = 0$ and $H_a : \alpha_p = 2/\sqrt(T)$ and $\beta_q = 2/\sqrt(T)$

For each scenario we will start with $p, q = 1$ and gradually increase until reaching either $p, q = log_2 T$ or $\sum_{p,q=1}^{\log_2 T} \alpha_p + \beta_q < 1$, or power in the range of 80-100%. We will also explore the negative parameter scenarios since they produce spectra with peaks different from low or high frequencies, thus more subtle to detect. We will do both the approximate theoretical formula calculation as well as a simulation study.

5.2.2 Theoretical power results

Tables 5.1 and 5.2 show the approximate theoretical power of the d_{00} test in Scenario 1 for AR and MA alternatives respectively. We notice that the power is not increasing or is even decreasing in AR(2) and AR(3) situations. This is due to the degree of the spectral polynom which results in a smaller d_{00} coefficient. As indicated in section 5.3 of the corrections, the very high frequency spectrum leads to the fact that the coefficient d_{10} will be higher and more meaningful to introduce a test, comprised of the three coefficients: $d_{0,0}, d_{1,0}, d_{1,1}$ and we will call it d_{11}. Similarly we can define d_{22} (which are the default settings of the General wavelet test(*genwwn*) using the three coarsest scales $j = 2, 1, 0$, resulting in seven coefficients to test.

Alternative	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Power of d_{00}	0.43	0.46	0.46	0.62	0.68

Table 5.1: Approximate theoretical power for d_{00} in Scenario 1: AR(p)

Alternative	MA(1)	MA(2)	MA(3)	MA(4)	MA(5)
Power of d_{00}	0.42	0.48	0.36	0.42	0.57

Table 5.2: Approximate theoretical power for d_{00} in Scenario 1: MA(q)

Alternative	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Power of *hwwn*	0.28	0.72	0.97	0.99	0.99

Table 5.3: Approximate theoretical power for *hwwn* in Scenario 1: AR(p)

Alternative	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Power of d_{11}	0.35	0.67	0.84	0.89	0.88

Table 5.4: Approximate theoretical power for d_{11} in Scenario 1: AR(p)

Results for d_{11} are reported in table 5.4 and they are better than d_{00}. Moreover, if we just go back to the *hwwn* test (which uses the total number of wavelet coefficients) reported in table 5.3, the power in the AR(3) case of Scenario 1 is 0.97 which is way higher than the global bases d_{00} and d_{11}. In this situation, this advantage has come at the cost of lower power of 0.28 in the AR(1) local alternative case. Furthermore, tables 5.4, 5.1 and 5.3 for d_{11}, d_{00} and *hwwn* tests show that the approximate power of d_{11} for Scenario 1 is quite better than d_{00}, but slightly worse than *hwwn*.

Test	Alternative	AR1	AR2	AR3	AR4	AR5	AR6	AR7
d_{00}	$T = 256$	0.44	0.4	0.15	0.13	0.2	0.18	0.11
	$T = 512$	0.44	0.41	0.16	0.15	0.22	0.21	0.13
	$T = 1024$	0.44	0.42	0.18	0.16	0.25	0.24	0.15
d_{11}	$T = 256$	0.34	0.41	0.54	0.53	0.32	0.18	0.13
	$T = 512$	0.33	0.43	0.59	0.58	0.36	0.22	0.15
	$T = 1024$	0.33	0.45	0.62	0.61	0.41	0.26	0.18
d_{22}	$T = 256$	0.27	0.33	0.46	0.46	0.35	0.42	0.53
	$T = 512$	0.25	0.33	0.49	0.5	0.43	0.54	0.65
	$T = 1024$	0.25	0.34	0.52	0.54	0.51	0.64	0.75

Table 5.5: Approximate theoretical power results for local Scenario 1a: AR(p)

Test	Alternative	MA1	MA2	MA3	MA4	MA5	MA6	MA7
d_{00}	$T = 256$	0.43	0.32	0.11	0.09	0.12	0.09	0.06
	$T = 512$	0.43	0.36	0.14	0.12	0.17	0.14	0.08
	$T = 1024$	0.43	0.38	0.16	0.14	0.21	0.19	0.11
d_{11}	$T = 256$	0.32	0.42	0.54	0.38	0.17	0.09	0.07
	$T = 512$	0.32	0.44	0.6	0.49	0.25	0.14	0.09
	$T = 1024$	0.32	0.46	0.64	0.56	0.33	0.2	0.13
d_{22}	$T = 256$	0.25	0.32	0.45	0.41	0.45	0.5	0.44
	$T = 512$	0.25	0.34	0.51	0.5	0.56	0.65	0.63
	$T = 1024$	0.24	0.35	0.55	0.56	0.64	0.75	0.77

Table 5.6: Approximate theoretical power results for local Scenario 2a: MA(q)

Tables 5.5 and 5.6 show the approximate theoretical power against emphnegative local alternatives scenario. The results in tables 5.5 and 5.6 were also validated by simulation and resulted in standard deviation 0.01 by doing 100 runs each consisting of 10^3 realizations for $T = 256, 512, 1024$. The conclusion is that for AR$(1)$to AR(2) we have that the largest coefficient in magnitude is $d_{0,0}$. For AR(3) to AR(5) the largest is $d_{1,0}$; and for AR(6) and larger, it is $d_{2,0}$. Jumping to figure 5.3, we can see the largest wavelet coefficients for those spectra. Tables 5.7 and 5.8 report the results for negative alternatives scenarios 1a and 2a with more observations for d_{22} test. The conclusion is that when the parameters are negative, detection is harder since we have mid-low or mid-high frequencies spectral peaks. In local alternatives scenario 3 and 3a (ARMA), we reach at least 80% power, even for $(1, 1)$ and $T = 2^7, 2^8, \ldots, 2^{12}$ by using the d_{22} test.

Test	Alternative	AR1	AR2	AR3	AR4	AR5	AR6	AR7	AR8
d_{22}	$T = 512$	0.25	0.33	0.49	0.5	0.43	0.54	0.65	0.65
	$T = 1024$	0.25	0.34	0.52	0.54	0.51	0.64	0.75	0.75
	$T = 2048$	0.24	0.35	0.55	0.58	0.59	0.73	0.82	0.82
	$T = 4096$	0.24	0.36	0.57	0.62	0.65	0.78	0.86	0.86

Table 5.7: Approximate theoretical power results for local Scenario 1a: AR(p)

Test	Alternative	MA1	MA2	MA3	MA4	MA5	MA6	MA7	MA8
d_{22}	$T = 512$	0.25	0.34	0.51	0.5	0.56	0.65	0.63	0.48
	$T = 1024$	0.24	0.35	0.55	0.56	0.64	0.75	0.77	0.67
	$T = 2048$	0.24	0.36	0.57	0.61	0.69	0.81	0.84	0.79
	$T = 4096$	0.24	0.37	0.59	0.64	0.72	0.84	0.88	0.85

Table 5.8: Approximate theoretical power results for local Scenario 1a: MA(q)

5.2.3 Spectrum estimation and periodicities

From response to correction points 1 and 4, the magnitude and sign of the wavelet coefficients of an ARMA spectrum give information regarding where the peak of the spectrum is located e.g. is it high or low frequency or in-between. However, this information alone is not complete and, as reported in the literature review of the thesis, wavelet thresholding is better used for curve estimation. In our situation, the largest and statistically significant wavelet coefficients could have been used in order to construct the spectral polynomial. For instance, figures 5.3 and 5.4 show three different spectra and their wavelet coefficients. While the Haar wavelets may result in higher power than, say, Daubechies's extremal phase with ten vanishing moments, using the latter results in better estimation of the wavelet coefficients that are zero, thus more suitable for spectrum curve estimation.

5.3 Nonlinear Models

The homogeneity of the variance is a strong assumption in the definition of white noise and in the performance of our tests as well as most white noise tests. Table 5.9 shows the empirical power results for 1000 realizations of the Time-modulated White Noise (*tmWN*) model such that: $X_t \sim N(0,1)$ for $t = 1, \ldots, 512$ and $X_t \sim N(0,2)$ for

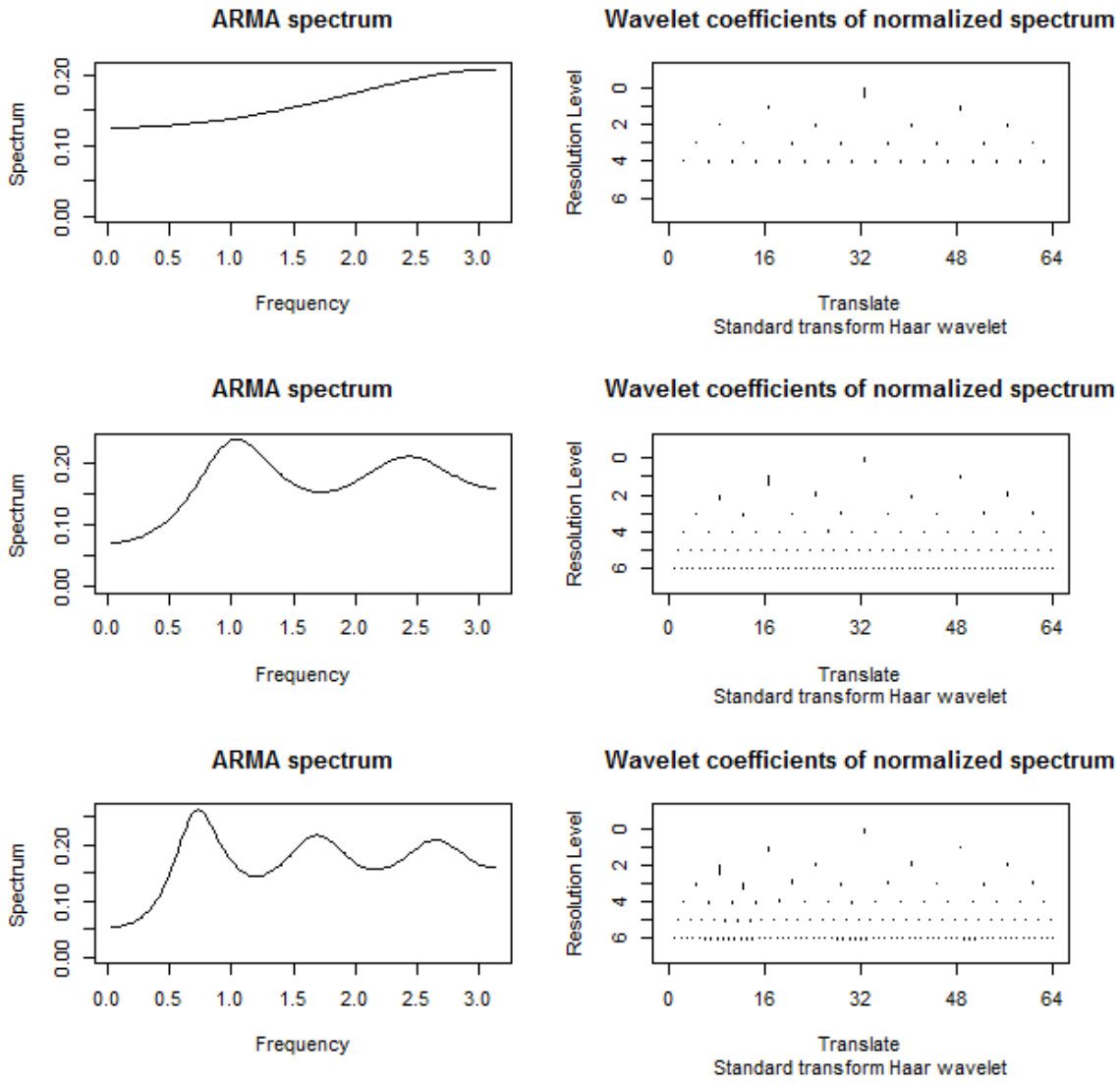

Figure 5.3: Spectra and their Haar wavelet coefficients for local alternatives Scenario 1a. Top: AR(1), middle: AR(4), bottom: AR(6).

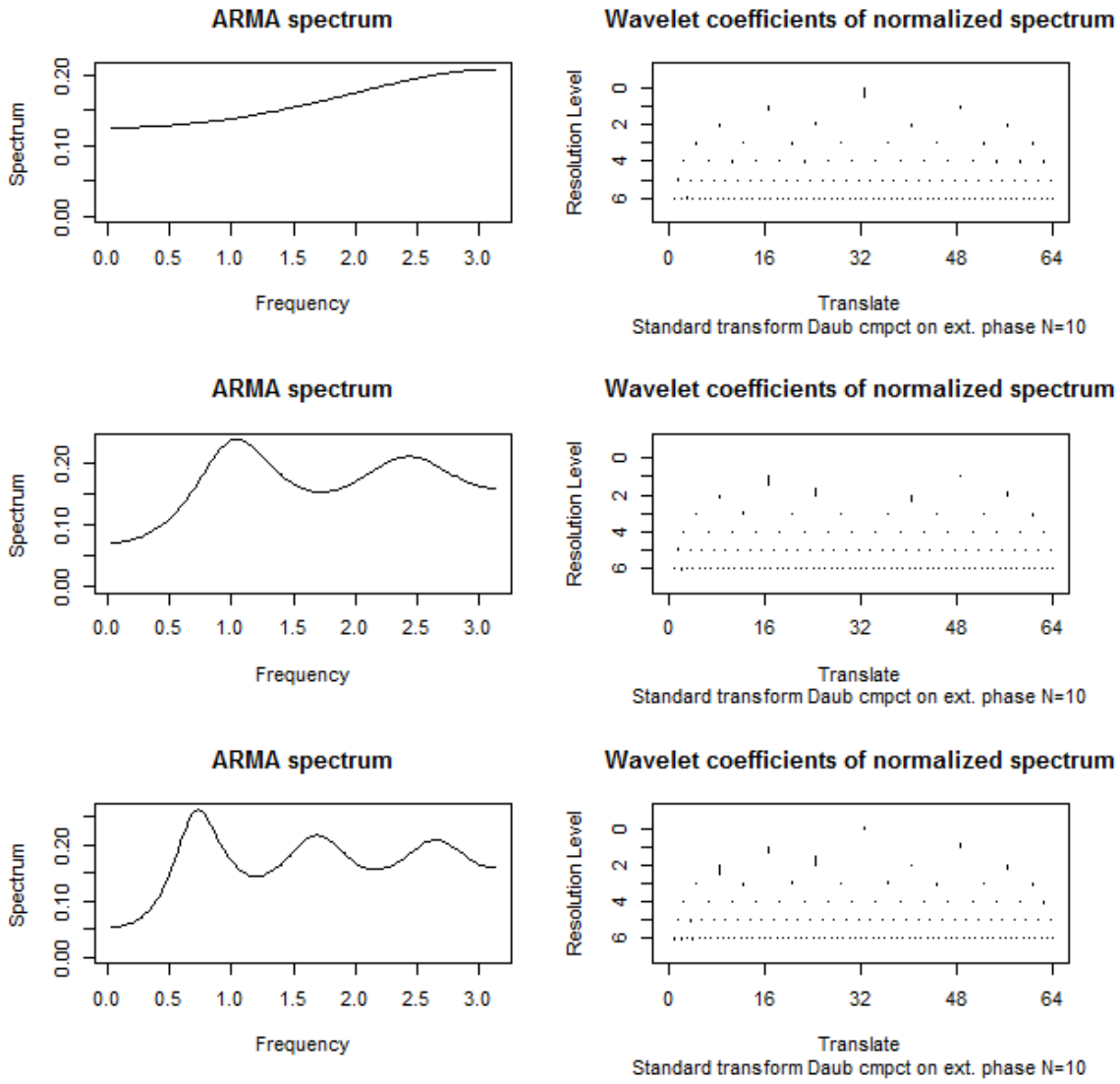

Figure 5.4: Spectra and their wavelet coefficients (10 vanishing moments) for local alternatives Scenario 1a. Top: AR(1), middle: AR(4), bottom: AR(6).

Test	hwwn	d00	genwwn_00	genwwn	genwwn_11	d_11	bartlett	box
Power	0.141	0.113	0.107	0.15	0.117	0.122	0.129	0.111

Table 5.9: Empirical power of various tests on 10^3 realizations following *tmWN* with $T = 1024$

$t = 512, \ldots, 1024$.

For GARCH$(1, 1)$ models, the situation is better. There are different options with respect to the combination of the parameteres. For computations we use the `R` package `fGarch` from Wuertz et al. (2016). Let $\epsilon_t \sim N(0, \sigma_t^2), \quad t = 1, \ldots, T$, the following parametrisation of GARCH is considered:

$$y_t = \sigma_t \epsilon_t, \tag{5.2}$$

where

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_p \epsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \\ &= \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2. \end{aligned} \tag{5.3}$$

The main parameters are α_i and β_i which determine the order of the model (p, q) similar to ARMA. We have selected six different combinations of parameter values for GARCH$(1, 1)$. The model parametrizations we test for are six with the convention $m = (\alpha, \beta)$ which gives: $m_1 = (0.3, 0.6)$, $m_2 = (0.05, 0.3)$, $m_3 = (0.05, 0.9)$, $m_4 = (0.1, 0.89)$, $m_5 = (0.4, 0.4)$, $m_6 = (0.89, 0.1)$, 10^3 realizations with $T = 1024$. Table 5.10 shows the results. We have added different versions of our wavelet tests, following the recommendations from section for more "global" wavelet basis to use in the tests. We notice that the *genwwn* test performs best in this study and the larger the quadratic serial dependence, governed by α_1, the better the statistical power. Furthermore, for financial applications, we know that the autocorrelation is evident when we consider the absolute values of the observations or their squared values. Doing 10^3 realizations with $T = 1024$ testing with

Model	hwwn	d00	genwwn_00	genwwn	genwwn_11	d_11	bartlett	box
m_1	0.15	0.25	0.25	0.37	0.34	0.32	0.33	0.24
m_2	0.06	0.08	0.08	0.07	0.06	0.08	0.07	0.08
m_3	0.07	0.08	0.09	0.1	0.1	0.1	0.09	0.08
m_4	0.2	0.15	0.15	0.28	0.22	0.21	0.21	0.15
m_5	0.15	0.25	0.21	0.33	0.34	0.32	0.33	0.28
m_6	0.51	0.58	0.56	0.77	0.76	0.75	0.73	0.59

Table 5.10: Empirical power results for various white noise tests against GARCH$(1,1)$ models m_1 to m_6: 10^3 relizations with $T = 1024$

genwwn gives the following results.

- For models — m_1, m_4, m_5 and m_6 we reach 0.99 power.
- For model m_2 we have powers of 0.13 with absolute values and 0.17 with squares respectively.
- For model m_3 we have powers of 0.42 with absolute values and 0.54 with squares respectively.

5.4 Simulation Study with low-magnitude parameters

We will investigate the case which is in-between local alternatives and "regular alternatives". As explained in chapter 4, Proposition 4 the general wavelet test *genwwn* test has a theoretical power function, which allows us to evaluate the statistical power of the test against a specific alternative hypothesis. Such procedures are common in medical statistics or design of experiments when one needs to calculate the needed number of observations to achieve a certain power of the test. A common specification in most areas is 5% threshold for type 1 error or statistical size, and 80% statistical power i.e. 0.2 probability for type 2 error. However, in practice, a trade-off may arise either:

1. when the number of observations is small
2. when the magnitude of the model parameter (assuming it is related to the test statistic) has a small magnitude, thus making itself hard to detect.

So, when thinking about ARIMA models, the latter case seems interesting. This is so, because when we have a large sample, we could detect very small efffects or parameters. On the other hand, when we have a moderate sample size (say range of $100-500$ observations), some small effects i.e. having an AR or MA parameter with magnitude $0.1-0.15$ may not be detected and the statistical power to be low, say, $15-50\%$. Furthermore, in the previous section, we saw that d_{00} test performs well in the scenario of $o(n^{-1/2})$ Guay et al, compared with their tests as well as a version of the Ljung-Box test, averaged over different lags.

We carried out a simulation study with the objective to see how our d00 and *genwwn* test would perform against Ljung-Box with different number of lags, 1 and 5. This aim is to help us for the ARH(p) verification which we are going to do via multiple testing procedure in chapter 7.

Next, we perform comparisons between the d00 test vs the Ljung-Box test in terms of empirical power for the AR(1) model i.e. $X_t = \rho_1 X_{t-1} + \varepsilon_t$ and MA(1) model $X_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t$ respectively, for the following parameter values for ρ_1 and θ_1: $0.1, 0.125, 0.15$. Furthermore, the crucial parameter for the Ljung-Box test is the lag. Usually, when we know the lag at which to look for dependence, this test works extremely well. However, selecting the lag p too large would decrease the empirical power of the test for smaller lag as shown in our simulation study in chapter 4. On the other hand, selecting p too small would miss dependence from higher lags, since the autocorrelations for the larger lags would not be included in the Ljung-Box test statistic. As we have illustrated in the previous section, for example, the application to the meteorological HADCRUT4 global dataset — there are situations when we do not know the exact lag beforehand.

So, Figures (5.5), (5.6) and (5.7) were created with 1000 replications for each scenario and 128, 256, 512 and 1024 observations for each replication. In all figures, the empirical power graph of Ljung-Box with lag 1 is with black, with lag 5 is red, d00.test with green and genwwn test with blue. We can also see that when the parameter and the number

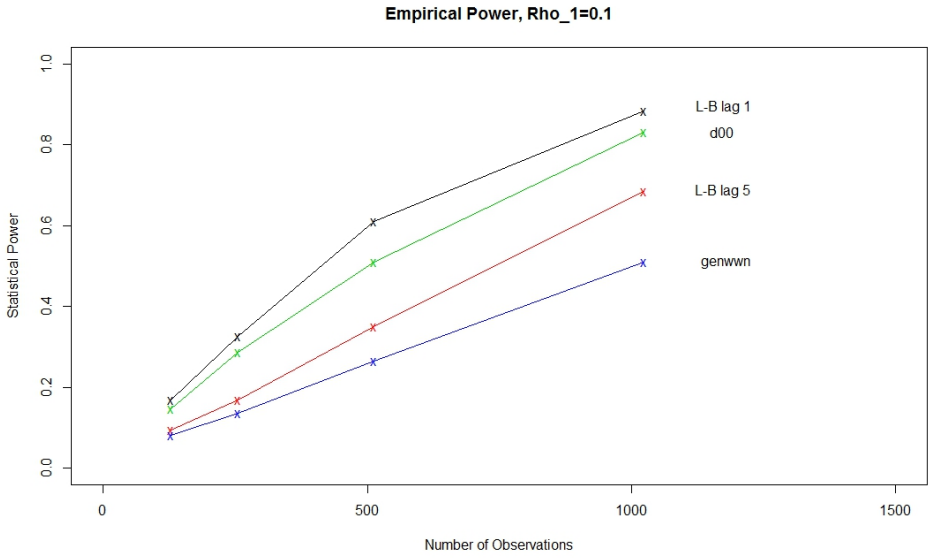

Figure 5.5: AR1 ($\rho_1 = 0.1$) Power

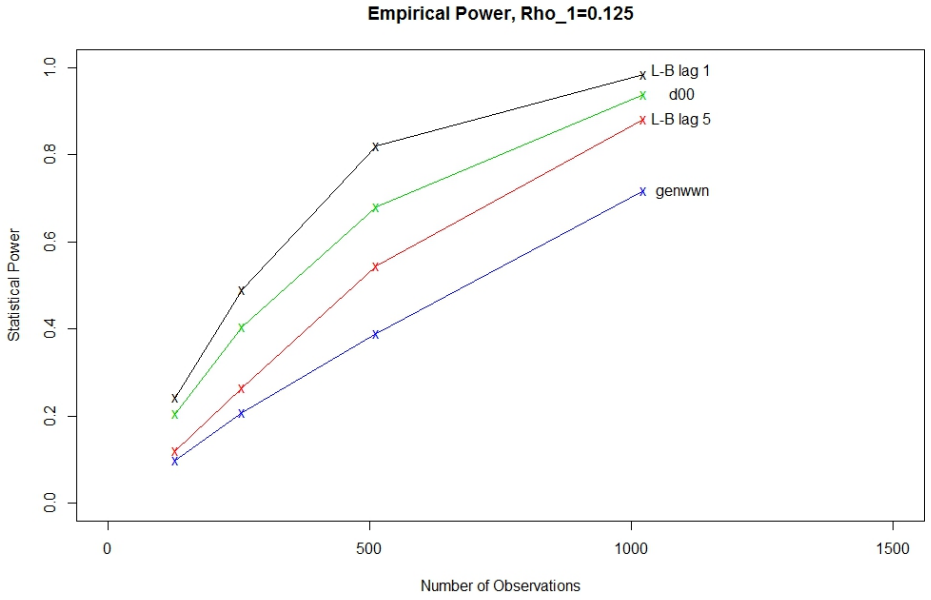

Figure 5.6: AR1 ($\rho_1 = 0.125$) Power

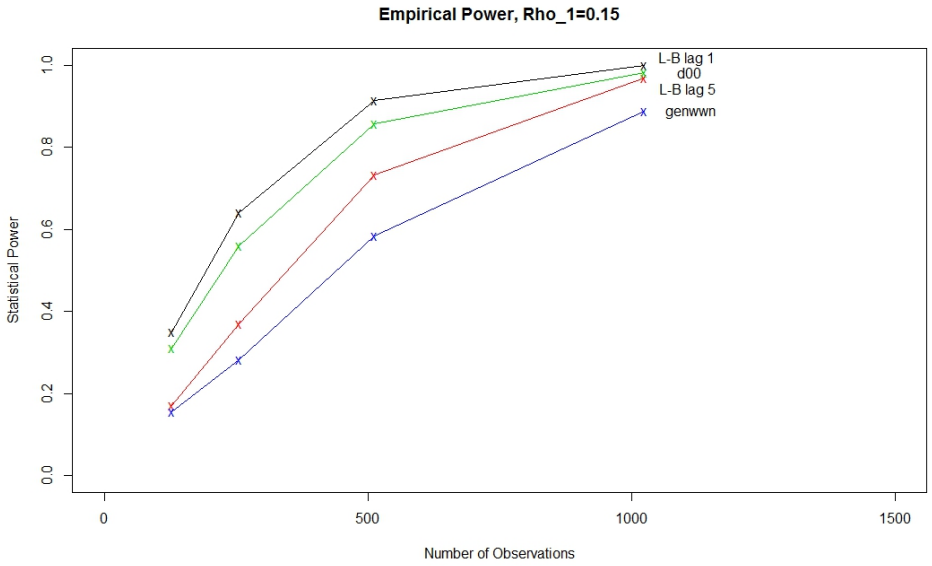

Figure 5.7: AR1 ($\rho_1 = 0.15$) Power

of observations increases then the statistical power increases. We performed the same analysis for the MA model, with similar results, shown in Figures (5.8) and (5.9)

5.4.1 What about moderate values of parameters

We also evaluated the performance when the autoregressive or moving average parameters are moderately large. In 5.10, we can notice, that even for 128 observations both the d00 and Ljnug-Box with lag 1 have over 80% empirical power.

5.5 Conclusion

So, based on those empirical findings for the statistical power, we may conclude that when we do not know the lag for the process we are analyzing, then it would be better to use the d00.test, because it has comparable power to Ljung-Box, when we know the lag, and better power when extra lags are included in Ljung-Box. Moreover, when we are faced with a multiple testing situation e.g. have a number of series and want to find their autoregressive or moving average order, the d00 might be helpful. The reason is in the number of comparisons that we need to perform when doing a multistage testing

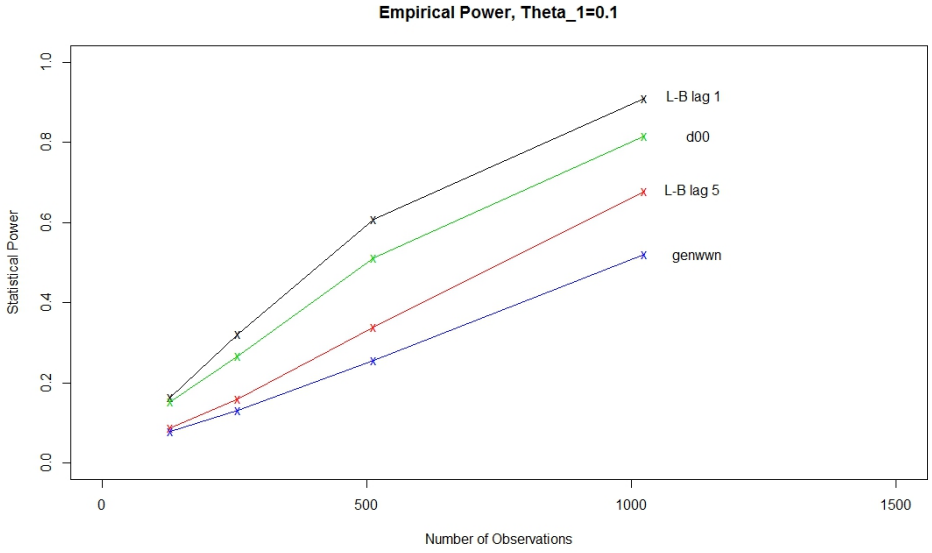

Figure 5.8: MA1 ($\theta_1 = 0.1$) Power

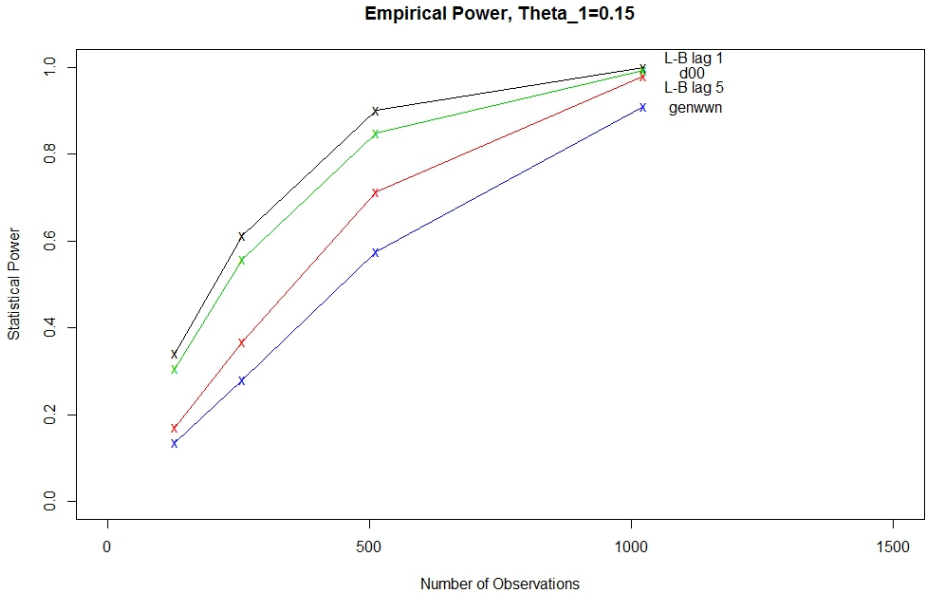

Figure 5.9: MA1 ($\theta_1 = 0.15$) Power

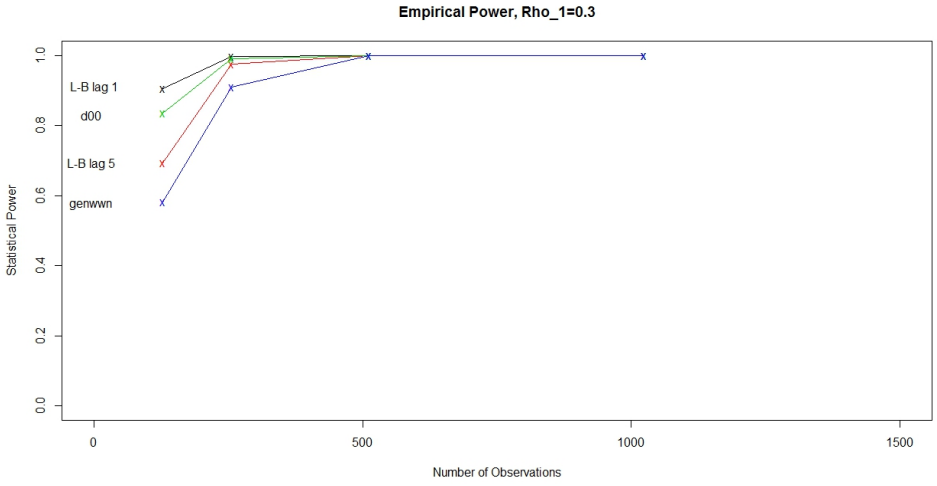

Figure 5.10: AR1 ($\rho_1 = 0.3$) Power

procedure as we will in chapter 7. For instance, if there is an AR(5) process where lag-5 parameter has the largest magnitude, we might first test it with low lags such as 2,3. On the other hand, if we start with lags 10 or 15, we might decrease the empirical power, if the process is of low order. This would be essentially important in case we would like to design an automated system for making those decisions. An infamous example is the performance of Google Flu Trends (GFT) whose forecasts were found very inaccurate from 2013 onwards, Lazer et al. (2014). Lazer et al. (2014) analyzed the issue and found that the aurocorrelations of the errors exceed the confdence interval up to lag 8 (Fig. S3 of supplementary material from Lazer et al. (2014)). An immediate question might be why not use AIC/BIC for univariate series? The answer is that we did a simulation study for order one univariate autoregressive/moving average process and we found out that, as long as the parameter is moderately large(> 0.2), using AIC of the `R ar` routine, it is correct 75% of the time. Thus, if want consistency of our test i.e. *the empirical power to go to* 1 *when the number of observations are going to infinity*, then d_{00} or Ljung-Box would be better since their power reaches 100% even for 256 observations.

Chapter 6

Two-dimensional Wavelet White Noise Tests

6.1 Introduction

In this chapter we are going to develop a two-dimensional wavelet test for white noise. We are primarily concerned with data that are recorded on a regular grid isomoprhic to an integer lattice. Typical examples of data include the yield of crops from a field or the colour-intensity of the pixels of an image. For example, Fig.6.1shows how two-dimensional white noise looks like on a grey scale. The statistical problem is the detection of spatial autocorrelation or trend. Similarly to univariate time series, when the data are white noise, then this is equivalent to a flat two-dimensional spectral surface. However, the situation in two-dimensions is more complicated than in one dimension, since there is more than one direction for the autocorrelation to take i.e. the lag has a direction.

6.2 Basic components for the two-dimensional test

This section defines the main components and their extensions from the univariate time series case to the spatial domain.

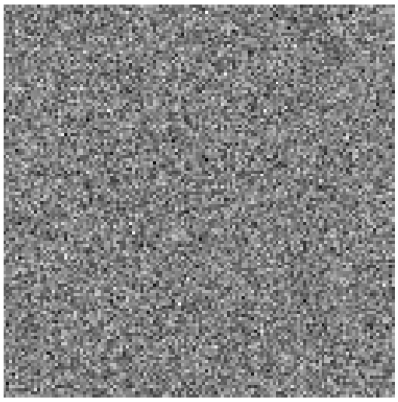

Figure 6.1: Two-dimensional 128×128 white noise as a grey-scale image

6.2.1 The 2D Periodogram

Starting with a square matrix of data $\mathbb{X}_{ts}$ for $t, s = 1 \dots T, \quad T \in \mathbb{N}$, we can calculate the 2D periodogram by the following:

$$I_{T,S}(\omega_1, \omega_2) = (2\pi)^{-2} T^{-2} \left| \sum_{t=1}^{T} \sum_{s=1}^{T} X_{t,s} e^{-i(t\omega_1 + s\omega_2)} \right|^2, \tag{6.1}$$

which can be computed at the Fourier frequencies $I_{p,q} = I_{T,S}(\omega_p, \omega_q)$, where $\omega_p = 2\pi p T^{-1}$ and $\omega_q = 2\pi q T^{-1}$ for $p, q = 1, \dots, T/2$ respectively.

6.2.2 Usage and properties of the two-dimensional periodogram

Borrowing from our work in one dimension in Chapter 3, we would like to use (6.1) for white noise testing. However, we need to be careful since, in two dimensions, things are more complicated. For example, if we start with an $n \times n$ matrix, then the two-dimensional periodogram at the positive frequencies would be of dimension $\frac{n}{2} \times n$. However, since we intend to use the two-dimensional Haar wavelet transform, we would need our input to be of dimension $n \times n$. Thus, we will need to replicate the periodogram *row-wise*. This

will be a replication of already available information, so it would not alter anything in this setup. To explain this further, let us rearrange the periodogram equation:

$$I_{T,S}(\omega_1, \omega_2) = (2\pi T)^{-2} \left| \sum_{t=1}^{T} \left[\sum_{s=1}^{T} X_{t,s} e^{-is\omega_2} \right] e^{-it\omega_1} \right|^2 , \tag{6.2}$$

The term in the square brackets represents the univariate discrete Fourier transform of the t-th row of the input data matrix. Thus the 2D discrete Fourier transform of a square matrix is computed as the univariate Fourier transform of each column from Cooley and Tukey (1965), while each row is replaced with its univariate Fourier transform. Consequently, similar to the one-dimensional case, the first row of the DFT matrix will contain the DC component. Then, there will be $n/2$ positive frequencies and, in the case of real data, the positive frequencies will be inversely repeated in rows $n/2 + 2$ to n. So our *row-wise* replication of the periodogram is essentially a rearrangement of the resulting two-dimensional discrete Fourier transform matrix.

6.2.3 The Theoretical Basis

The result for the univariate periodogram ordinates' distribution, from Brockwell and Davis (1991) page 344, applies to the two-dimensional case. The asymptotic probability distribution of the two-dimensional discrete Fourier transform, of an independent and identically distributed bivariate random vector (at the Fourier frequencies, apart from $\frac{1}{2}$), is derived in Brilliger (2001), Theorem $4.4.1$ — as a complex Gaussian distribution. Therefore, the squared magnitude of a standard complex Gaussian random variable $z = X + iY$, where $X, Y \in \mathbb{R}^k$ will be Exponentially distributed with parameter $\lambda = 2\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimated variance of the input data, taken as a univariate series. Thus, we will model the two-dimensional periodogram ordinates as independent and identically distributed Exponential random variables. Moreover, the same result is used in Pawitan (1996) for deriving an automatic method for cross-spectrum estimation of a bivariate time

series. Another example, where this property has been used in two dimensions, is Rao et al. (2014).

6.2.4 The 2D Haar Wavelet Transform

We are going to use the 2D decimated (orthogonal) Haar wavelet transform as described in section 2.6.7 and Nason (2008) page 76. However, unlike the univariate transform where there are only detail-level and scaling function coefficients, in two-dimensions we are going to have four sets of wavelet coefficients — horizontal, vertical and diagonal (detailed) — and scaling function coefficients. In the next section we will derive their distribution under the null hypothesis of $\mathbb{X}_{ts}$ being white noise.

6.3 Distribution of the $2D$ Haar wavelet cofficients

Proposition 5. *Assuming an Exponential(λ) distribution for the two-dimensional periodogram ordinates from section 6.2.3, then for a general scale $l = J - j$, the characteristic function of the diagonal coefficients will be:*

$$\phi_{Wd_{[l]}} = \frac{\lambda^{[2^{2l}]}}{(\lambda^2 + t^2/4)^{[2^{2l-1}]}} \tag{6.3}$$

with density function given by :

$$g_m(x) = \frac{\sqrt{2m}\exp(-\sqrt{2m}|x|)}{2^{2m-1}(m-1)!} \sum_{j=1}^{m} \frac{(m+j-2)!}{(m-j)!(j-1)!} \left(2\sqrt{2m}|x|\right)^{m-j}, \tag{6.4}$$

where $m = 2^{[2(J-j)-1]}$ for $j = 0, \ldots, J-1$ and, we assume that $\sigma^2 = 1$ i.e. $\lambda = 2$.

Proof: See Appendix A.5.

The difference with the univariate *HWWN* test is only in the scaling m parameter for a given scale l i.e. $m_{univariate} = 2^{[(J-j-1]}$ for $j = 0, \ldots, J-1$ and $\sigma^2 = 1$. Similarly, when $m = 1$, the distribution is the Laplace distribution. It is probably interesting to mention that a form of this distribution, using the Macdonald/Bessel function of the second kind, has been derived in Fisher (1915) as a the limiting distribution of the correlation coefficient of a bivariate Gaussian random vector.

6.3.1 Empirical distribution of the $2D$ wavelet coefficients

Similar to the details outlined in section 4.5, the convergence of the wavelet coefficients to the limiting Macdonald distribution is good enough for our implementation of the wavelet tests. Furthermore, in the $2D$ case, the parameter m of the Macdonald distribution is growing faster than the $1D$ since $m_{1D} = 2^{J-j-1}$ and $m_{2D} = 2^{2(J-j)-1}$, therefore the convergence of the Macdonald to the Normal distribution. For example, for the finest scale $j = J-1$, the parameter $m_{1D} = 1$, whereas $m_{2D} = 2$; then for $j = J-2$: $m_{1D} = 2$ and $m_{2D} = 8$, and so on. Figure 6.2 shows the empirical distribution of the finest-scale diagonal wavelet coefficients against their theoretical Macdonald curve with $m = 2$ and dimensions 16×16, 32×32, 64×64 and 128×128 for 1000 realizations. As we have seen in the $1D$ case, for coarser scales, the distribution would be closer and closer to the standard Normal. The situation for the horizontal and vertical coefficients is similar. For example, the top left graph of figure 6.2 shows the 64 diagonal coefficients corresponding to $T = 256$ (16×16) — if we compare with the top right picture of figure 4.4, which shows 1000 realizations of 64 Gaussian observations using `rnorm` in `R`, the convergence of the empirical wavelet coefficients to the Macdonald seems to be no worse than the one for the Gaussian.

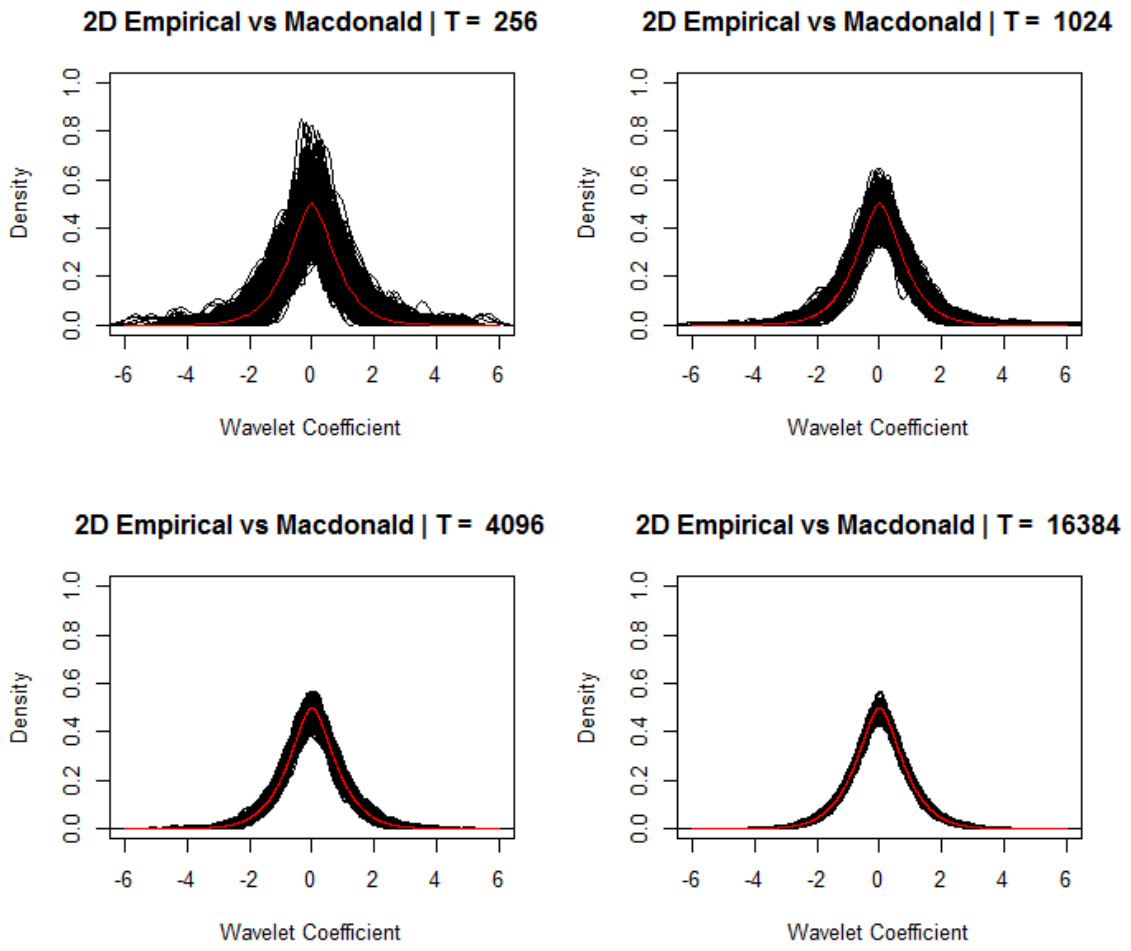

Figure 6.2: Top left to bottom right respectively: black — 1000 realizations of Empirical distribution of finest-scale diagonal $2D$ Haar wavelet coefficients of 10^3 realizations from Gaussian white noise with $T = 256, 1024, 4096, 16384$, red — theoretical Macdonald curve

.

6.4 Two-dimensional *HWWN* test procedure

In this section we propose the algorithm for two-dimensional wavelet white noise testing. We will also mention some computational details that need to be considered at each step.

Suppose we have a square matrix of data $\mathbb{X}_{ts} = \{x_{ts}\}$ of dimension $n \times n$ i.e. $t, s = 1 \dots n$. Let $\bar{x} = \frac{1}{n^2} \sum_{i,j=1}^{n} x_{ij}$ be the global mean of the data and we denote the estimated variance of $\mathbb{X}_{ts}$ (considered as univariate series for the moment) by $\hat{\sigma}^2$, where $\hat{\sigma}^2 = \frac{1}{n^2-1} \sum_{i,j=1}^{n} (x_{ij} - \bar{x})^2$.

The 2D *HWWN* algorithm is:

1. Standardize the data $\mathbb{X}_{ts}$ by subtracting the mean $\bar{x}$ from each x_{ts} and dividing each x_{ts} by the estimated standard deviation (i.e. $(\hat{\sigma}^2)^{-1/2}$). In `R` we use the `scale` function.

2. Form the two-dimensional periodogram of $\mathbb{X}_{ts}$, take the positive frequencies, replicate them row-wise and execute the 2D wavelet transform on the resulting $n \times n$ matrix.

3. Obtain ordinary p-values for all of the wavelet coefficients according to the their cumulative distribution functions (cdf). The distribution function can be calculated numerically from the density function — for which we already have the analytical form from (6.4).

4. Use Bonferroni correction or other multiple comparisons' adjustment technique.

5. Take the minimum value of the resulting list of adjusted p-values and assign it as the p-value of the test.

This procedure will be used for our simulations later.

6.5 Spatial Statistics White Noise Tests

This section briefly reviews white noise testing in spatial statistics. We will focus on two of the most popular tests: Moran's I and Geary's C.

In order to conform with other sources, we will change notation here. The idea is that the usual spatial lattice has regions that can be numerated from 1 to their count. Our data matrix $\mathbb{X}$ of dimensions $n \times n$ will be stacked into a $n^2 \times 1$ vector $\mathbf{y}$, let also $N = n^2$, thus $\mathbf{y} = (x_{11}, x_{12}, \ldots, x_{1n}, x_{21}, x_{22}, \ldots, x_{nn}) = (y_1, y_2, \ldots, y_N)^T$

6.5.1 Popular Measures of Spatial Autocorrelation

A popular spatial autocorrelation measure is Moran's coefficient for autocorrelation from Moran (1950). Its test statistic is:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6.5}$$

where $N = n^2$, y_i are the observations i.e. $\mathbf{y} = (y_1.y_2, \ldots, y_N)^T$, $\bar{y}$ is the mean of $\mathbf{y}$ and $W = w_{ij}$ is a $N \times N$ matrix of spatial weights such that $w_{ii} = 0$. The idea is that, when evaluating the spatial autocorrelation, Moran's I would include the neighbouring observations with different weights, for example, but not the observation itself.

Under the null hypothesis of two-dimensional Gaussian white noise, $E(I) = -\frac{1}{N-1}$, from Moran (1950), which is close to zero when N is large. As with Pearson's product-moment correlation, Moran's I varies from -1 to 1. Similarly to regular correlation, positive values indicate positive autocorrelation and negative values indicate negative autocorrelation. Thus, in order for the significance to be evaluated, one needs to construct a test statistic that studies the deviation of this expected value. The usual test, Cliff and Ord (1972), treats the empirical value of I as a standard normal deviate (under the null hypothesis of Gaussian white noise) and tests it via the usual normalisation — the test

statistic, say, K would have the following form: $K = \frac{I - E(I)}{\sqrt{Var(I)}}$

Another measure is Geary's C from Geary (1954) is defined as:

$$C = \frac{(N-1)}{2\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6.6}$$

where N is the number of spatial units, $\bar{y}$ is the mean and W_{ij} is a matrix of spatial weights. It is also known as the contiguity ratio. Geary's C varies from zero to two, with the value one indicating no spatial autocorrelation, values from zero to one indicating positive autocorrelation and values from one to two indicating negative autocorrelation.

The measure from Moran seems to behave like a global one, whereas Geary's is more local. This is so, because, after close inspection of their formulae, one can notice that the numerator in Moran's I is the *weighted* variance of all the n^2 datapoints, whereas in Geary's C, the numerator expresses the deviations among the possible pairs of observations themselves rather than deviations from their mean value.

6.5.2 Brief literature review of spatial autocorrelation tests

We will concentrate on Moran's coefficient (6.5), since it has been popular in the spatial statistics literature, and examine its relations to standard time series methods.

One of the first extensions of Moran's test for spatial autocorrelation among residuals, meaning to test the null hypothesis that it is zero, is developed in Cliff and Ord (1972). This test is essentially based on one of the early tests for time series residual autocorrelation — Durbin and Watson (1950), later refined by Durbin and Watson (1971). A procedure, which is suited for estimating parameter of spatial autoregression, is provided in Ord (1975). Eventually, Burridge (1980) shows that the test from Cliff and Ord (1972) is a special case of a Lagrange multiplier test, based on a more general mathematical op-

timization paradigm from Silvey (1959). Similarly to other time series portmanteau tests, the limiting distribution is Chi-squared. More recently, Robinson and Rossi (2014) develop an improved Lagrange multiplier test for small spatial datasets in the autoregression setting. Another approach is taken by Li et al. (2007) which uses the spatial autoregressive model and profile likelihood estimator to show that when (6.5) is close to 1, another estimator is needed. However, they conclude that (6.5) is a good estimator when the true autocorrelation coefficient is close to zero.

6.5.3 An illustrative example: the wheat data

In this section we will analyse on one of the classical spatial statistics datasets: the wheat data from Mercer and Hall (1911). This dataset is extensively used for examples in Cressie (1993). The dataset has also inspired interesting discussions among generations of statisticians. For example, the following quote from Cressie (1993), page 457 considering the models for spatial autoregressive Gaussian model of Whittle (1954), says:
"*In my opinion, Whittle's conclusion about the poor fit of spatial models to the Mercer and Hall data is less due to the inconsistent likelihood approximation chosen and more due to his not accounting for the large scale variation (trend) in the data*"

Description of the Data and its Issues

We will now explain the dataset and its issues in detail. The agricultural experiment considers the yields (in pounds) from wheat grown on a certain region of fields at Rothamsted's Experimental Station in England (where R.A. Fisher also worked). The region for the experiment consists of a 20×25 grid (lattice) of plots (500 in total), with 25 coordinates running East to West and 20 running from North to South. The object of the study was determining the plot size that would "*reduce the inevitable error within working limits*", Cressie (1993), chapter 7. The conclusion of Mercer and Hall (1911) was that 5 plots or $1/40$ of an acre should give adequate precision. "*A uniform area of* 1 *acre was har-*

vested in separate plots, each $1/500$ *acre in area. The wheat sheaves were then threshed out by hand and weighed*", Cressie (1993). Furthermore, Mercer and Hall (1911) say that "*since the [frequency] curve fits the [Gaussian] one as well as may be expected ... we may conclude that the material is fairly homogeneous*" However, Cressie (1993), chapter 4 undertakes exploratory analysis in detail with a *median-polish* procedure and shows that the Gaussian assertion was not plausible because there was a trend in the east-west direction. Moreover, Besag (1974) and Besag (1977) say that their model fits were also not good. Afterwards, Cressie (1985) shows that there is a non-linear trend in the east-west direction.

Due to the ambiguity regarding the plot size, there have been some variations in the literature — for example Whittle (1954) and Besag (1974) use 10.82×11 ft whereas Ripley (1981) uses a square with 11ft sides. For our analysis we will use the plot dimensions in metres used in Cressie (1993), chapter 4, of $3.30m$ (north-south) $\times 2.51m$ (east-west). They will serve as local latitude and longitude respectively i.e. we have 500 wheat yield measurements in total 20×25 plots, each $3.3 \times 2.51m$ which form our grid.

Cressie's Median-polish analysis of the data

We will follow Cressie (1993), chapter 4 for the analysis of the data. For the algorithmic details of the median-polish procedure, we refer to Cressie (1993), chapter 3 or the originator of the procedure — Tukey (1977). Essentially, this procedure is similar to two-way analysis of variance (ANOVA) in the regard that it tries to estimate a *median* total effect (rather than *mean*), and column and row effects of the grid respectively. The difference with the ANOVA is that median-polish can be thought as of minimising the modulus of the errors rather than the square.

To summarize, Cressie (1993), page 250 discusses the possible and previous analyses of those data with respect to the already mentioned trend. A total surface of the data with

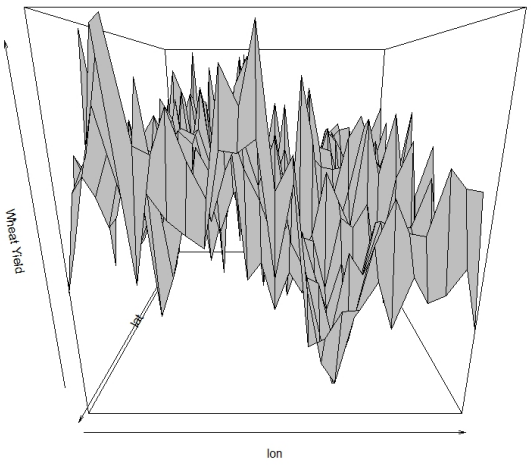

Figure 6.3: Wheat Raw Data Total Surface

respect to the local latitude and longitude can be seen in Fig. 6.3. We can notice that some kind of trend appears in the longitude direction.

Ripley (1981) found a spectral peak in the east-west direction which confirms what is visible from the figure. Now, for comparison, let us have a look at the surface of the median-polish residuals in Fig. 6.4 — we notice that the trend has been considerably reduced, though there is still some irregularity left.

Our Analysis of the Data

Since our wavelet test requires a square and dyadic input, we decided to select 256 data-points our of the 500, based on their coordinates. We selected a set of datapoints and we already know it contains a suspected trend — out of the 20×25 regions grid, we selected $4:19 \times 4:19$, total of 256 observations.

Firstly, let us simply test the selected subset of raw data with our test — the p-value is 0.026, so we have evidence to reject the null hypothesis of spatial white noise. Secondly, let us examine the raw data surface in Fig. 6.5 — we can notice the sharp direction of

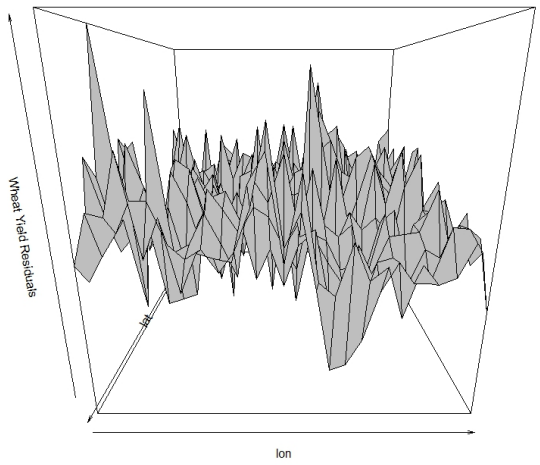

Figure 6.4: Wheat Data — Residuals Surface after Median-Polish

the trend. Next, let us do the median-polish decomposition on the selected subset — on Fig. 6.6 we can notice the already-discussed trend's disappearance with some remaining effects on the right. Eventually, we also did the test on the median polish residuals of the 256-datapoints subset (first using `medpolish` from `stats` package) and the p-value was 0.227. Therefore, our conclusion is that, after removing the trend, the process is white noise.

Furthermore, for illustration, we can use the `image` function in `R`, which can be seen on Fig.6.7 — the reddish nuance is quite clear in the upper right corner of the left image, which are the raw 256 datapoints. For comparison, Fig.6.8 is a realization of Gaussian white noise, plotted with the `image` function in `R`.

As part of the analysis, Cressie (1993) also looks at a graph of the column effects only (which correspond to the east-west direction trend). On Fig. 6.9, we can check that in our subset they are largely preserved.

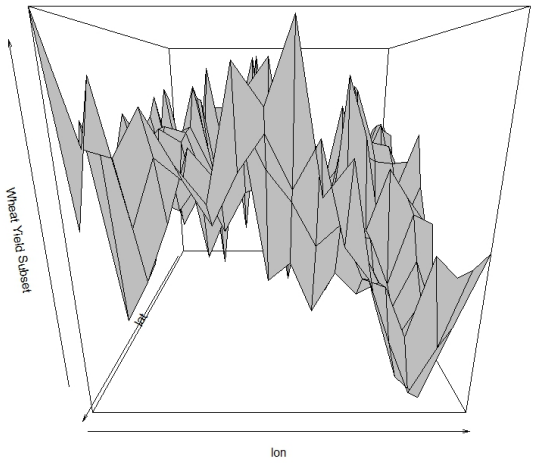

Figure 6.5: 256 Wheat datapoints subset's surface — raw data

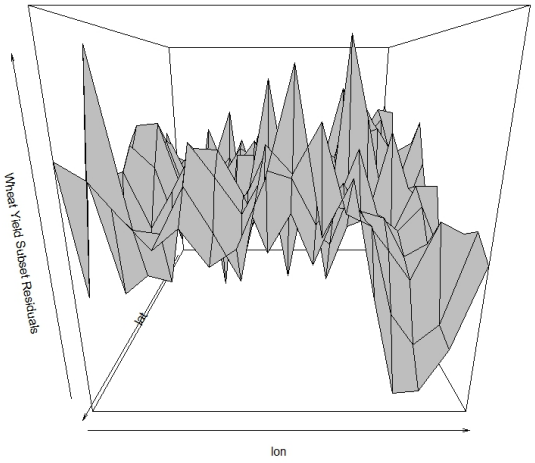

Figure 6.6: 256 Wheat datapoints subset's surface — residuals after *median-polish*

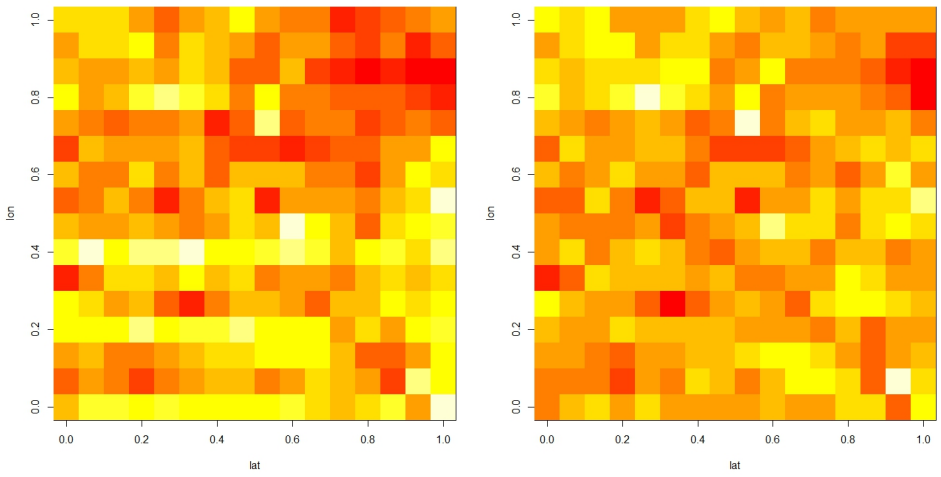

Figure 6.7: Wheat 256 datapoints subset's images — total (left) and residuals after median polish(right)

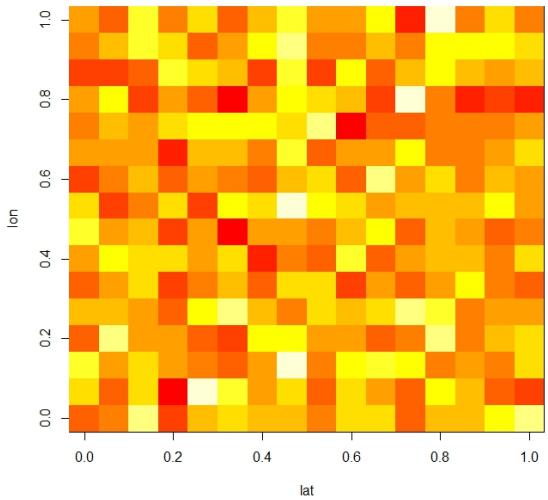

Figure 6.8: Image of 256 observations of Gaussian white noise

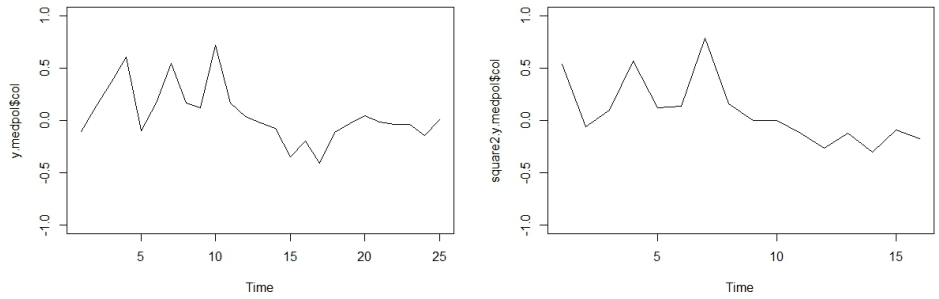

Figure 6.9: Total column effects vs plot number (east-west) (left) and the same for our 256-datapoint subset(right)

Augmenting the wheat data

Since we have 500 wheat yields, we will expnad them to $1024 = 32 \times 32$ by 'fill-in' the data with Gaussian noise bearing the same mean value and standard deviation as the original data. For the wheat data, the average yield $\bar{y} = 3.95$ with standard deviation $\sigma_y = 0.46$. So, we will take a 32×32 matrix $\mathbf{Y} = [y_{ij}]$ with values $\{y_{ij}\} \sim N(3.95, \sigma_y = 0.46)$, $i, j = 1, \ldots, 32$ and then plug the 20×25 wheat data inside $\mathbf{Y}$. If $\mathbf{X} = [x_{mn}]$, $m = 1, \ldots, 20$, $n = 1, \ldots, 25$ are the wheat data, then by putting $\mathbf{Y_{7:26,4:28}} := [x_{mn}]$, the matrix $\mathbf{Y}$ is to be the augmented wheat data. Figure 6.10 shows the original 20×25 wheat data on the left and the augmented wheat data on the right. Despite the increased resolution, the downward trend is still visible (the dark nuance spot around of the original data at coordinates $(0.8, 0.6)$ versus the augmented data at $(0.7, 0.6)$). Performing both the 2D HWWN test and Moran's I on the augmented data results in rejecting the null hypothesis.

In summary, the dyadic restrictions might be meaningful to low-resolution data. A way to by-pass the restrictions would be to pad with zeros or mean value and then discard those coefficients. An implementation of non-dyadic wavelet transform from Whitcher (2015) could be used to implement such tests.

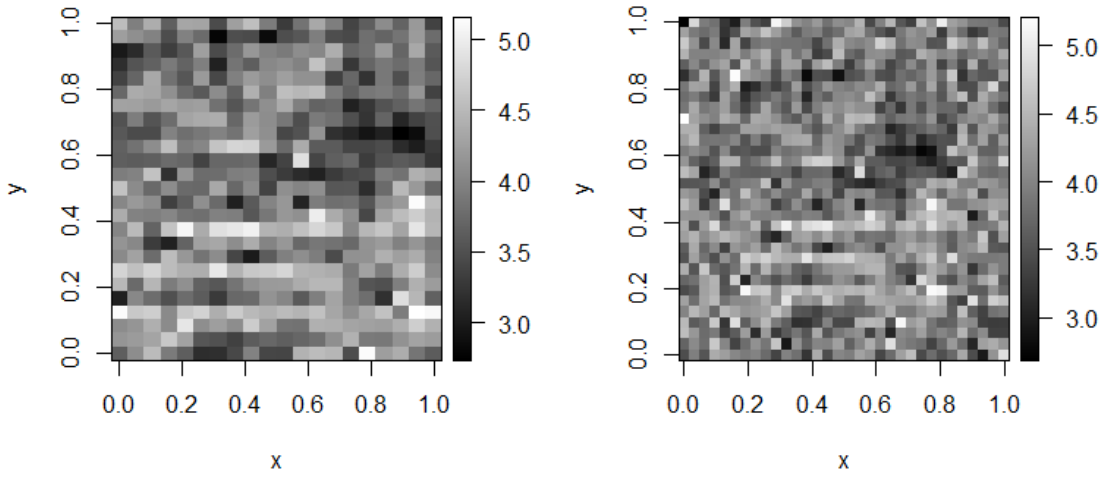

Figure 6.10: Left: image of the original 20×25 wheat data. Right: image of the augmented 32×32 wheat data.

6.5.4 Software for Spatial Statistics

One of the popular and comprehensive R packages for spatial statistics is `spdep` from Bivand and Piras (2015); Bivand et al. (2013). The package can define spatial weights — via a matrix, via a list and via neighbours, which are special kind of objects for spatial statistics. Furthermore, there are functions for testing the significance of Moran's I, Geary's C as well as a range of similar tests based on *Lagrange multipliers* as in Silvey (1959). Moreover, the functions also support permutation tests for the mentioned statistical tests, which are applicable when we are not sure of for the distribution under the null hypothesis of spatial white noise.

6.6 Spatial Autocorrelation Testing

6.6.1 The Matrix of Spatial Weights

At first sight, it might not seem appealing why our 2D wavelet test should be used for spatial autocorrelation. However, examining closely equation (6.5), one notices that the crucial bit is the matrix of weights $W = W_{ij}$. Because spatial statistics is primarily driven by applications (e.g. geostatistics, petroleum drilling, map coordinates), this means that there are different weighting schemes in existence. For example, the first and logical thing to do (for measuring something on a spatial grid) would be to design the weights properly. For example, having an $n \otimes n$ matrix and setting weights of $\frac{1}{j}$ for each spatial unit e.g. starting from position $(2, 1)$ for $j = 2 \dots N$ would mean that the farther the location, the lesser the weight e.g. heat or disease maps. Essentially, this represents the Euclidian distance used for weighting of the points when calculating the Moran's spatial autocorrelation from equation (6.5).

Moreover, for some problems, we could be more stringent and want only the immediate neighbours of an entry to have a non-zero weight. Furthermore, there is also the issue for boundary conditions i.e. if we have a 16 by 16 dataset and we want (in the calculation

of eq. (6.5)) every entry to be weighted with its nearest neighbours (being 4 — left, right, above, below). This means that all entries in the first row do not have a neighbour above. The $(1,1)$ entry has only 2 neighbours (right and below) and so on. So, for computation in the *spdep* package, there is the option to choose type of weights and boundary conditions — rook (based on contiguity, a common boundary) and queen (also common corners), and torus — the last meaning that we can fold our lattice/grid which will ease and unify computation for all entries. This is similar (but not exactly the same) as the wavelet boundary conditions options in *wavethresh* where the user can select periodic or symmetric boundary extension.

Furthermore, let us illustrate how an arbitrary weighting matrix would look like when we want to have nearest neighbour contiguity:

$$W_{N,N} = \begin{pmatrix} 0 & w & w & 0 & 0 & \cdots & 0 & w & w & \cdots \\ 0 & 0 & w & w & 0 & \cdots & w & w & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ w & 0 & w & 0 & 0 & \cdots & w & w & 0 & \cdots \end{pmatrix} \tag{6.7}$$

We can notice that (6.7) is a sparse matrix — there are four non-zero entries in each row denoted by w - a real number, but the matrix is of dimensions $N \times N$. When we want to have rook contiguity with torus linkage, this means that every observation's attribution to correlation (i.e. in the numerator of equation (6.5)) is weighted together with its four immediate neighbours. The torus means that our lattice can be folded. For instance, the (1,1) entry of an input matrix $\mathbb{X}_{n,n}$ has the entries $(1,2), (2,1), (1,n)$ and $(n,1)$ as immediate neighbours. Say, we have an input of 8×8 square matrix. This means that $N = n^2 = 64$, $\mathbf{y}$ is 64×1 vector and W is 64×64. The first row of W would have four non-zero entries — they would be precisely in positions $(2), (8), (9)$ and (57) which correspond to $(2,1), (8,1), (1,2)$ and $(1,8)$. The first step in calculating the spatial autocorrelation from equation (6.5) would be to multiply — $W \times \mathbf{y}$

6.6.2 Weighting and the 2D Haar Wavelet Transform

However, if we look closely at the 2D Haar wavelet coefficients formation, we can see that, depending on the scale, different regions of the input matrix participate in the detailed coefficients calculation (with different observation blocks). Thus, a direct comparison of our test with any modification of Moran's test might be hard, though not impossible. For instance, if we consider only the finest scale detailed Haar coefficients, we can check that we are analyzing our lattice in orthogonal fourplets (though in spectral domain, rather than time) which corresponds to having four neighbours; on the next scale that would be 16, next 64 and so on.. On the other hand, we could think of the wavelet approach as a generalization of several tests with different weighting schemes. This is so because we are working in the frequency domain, thus violation of uniformity for any (two or more) frequency bands (e.g. contrast between high and low frequencies across locations) would be caught although, it may not correspond to a specific weighting scheme. For example, we could try doing the wavelet test using only finest scale coefficients or coarsest scale for that matter.

From our univariate work in chapter 4, we have seen that the approximate distribution of the wavelet coefficients of Gaussian white noise performs well in comparison with other tests. In the 2D case, the limiting distribution of the Haar wavelet coefficients is the same, but even closer to Gaussian. This situation makes it worthwhile to pursue statistical testing. Furthermore, in two dimensions we have three types of wavelet coefficients to use as a proxy for testing constancy of the spectrum. For example, figures 6.11 and 6.12 show the periodograms of a spatial autoregressive model(SAR) model with parameter $\rho = -0.8$ from section 6.7.1 of the thesis for both low and high frequency 2D spectrum. There is notable difference in the shapes of the surfaces. Moreover, if we look at the images of the 2D periodograms, the contrast is even more striking — shown on figure 6.13.

In similar setting, the 2D test might be used for analysis of images i.e. to detect if there is a non-random pattern emerging.

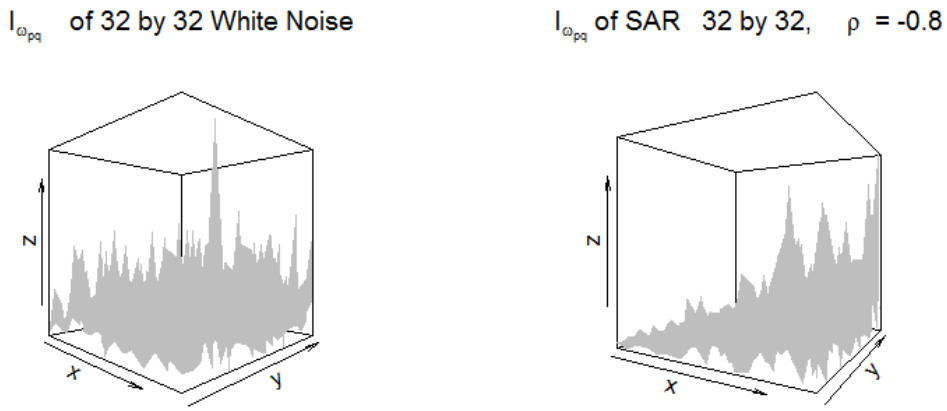

Figure 6.11: Left: 2D periodogram of 32×32 Gaussian white noise. Right: 2D periodogram of 32×32 SAR model.

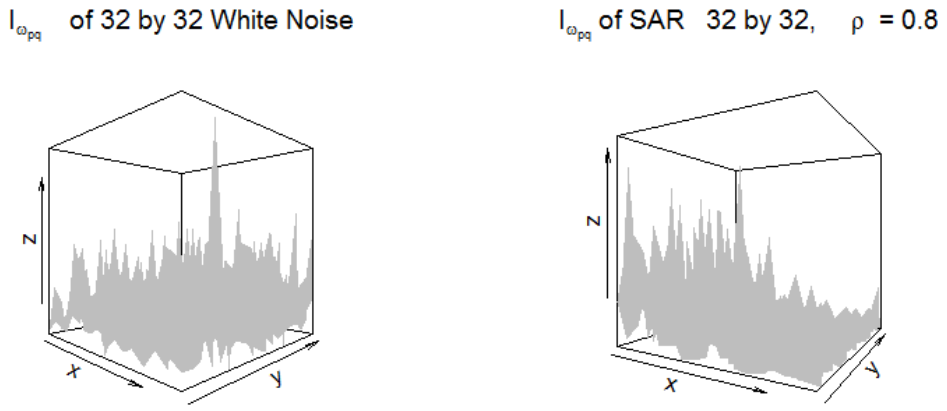

Figure 6.12: Left: 2D periodogram of 32×32 Gaussian white noise. Right: 2D periodogram of 32×32 SAR model.

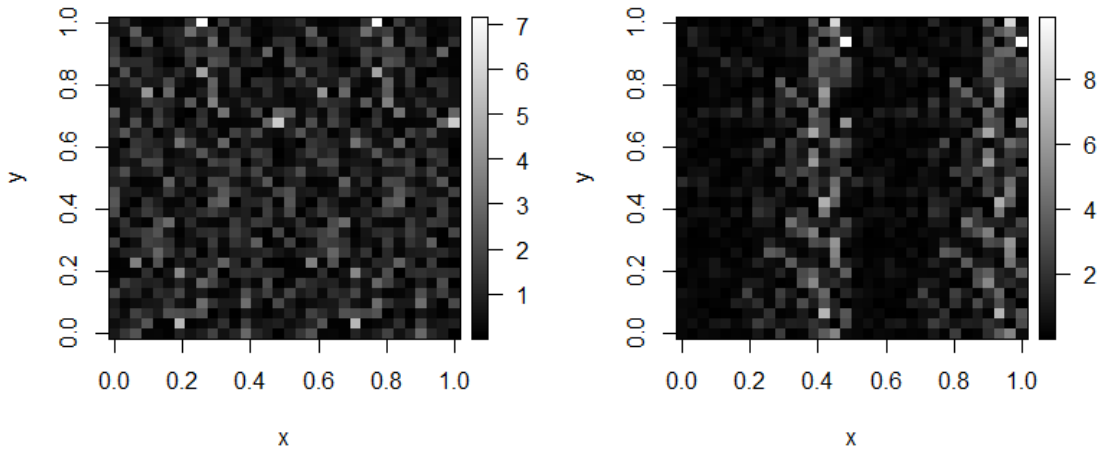

Figure 6.13: Left: image of the 2D periodogram of 32×32 Gaussian white noise. Right: image of the 2D periodogram of 32×32 SAR model ($\rho = -0.8$).

6.7 Simulation Results

In this section we will compare the 2D Haar wavelet test procedure *HWWN* (as described in section 6.4) with the Moran's test from (6.5) on spatial white noise and autoregressive process of order one with the procedures from the R package `{spdep}`.

The next section reviews spatial autoregressive models since as they differ from their univariate time-domain counterparts.

6.7.1 The Spatial Autoregressive Model (SAR)

A direct extension of the univariate autoregressive process(UAR): $x_t = \rho x_{t-1} + \epsilon_t$, where $|\rho| < 1$ and $\epsilon \sim N(0, \sigma^2)$ to spatial domain is possible, however, it must be done with caution. This is because, in the time domain, time only flows in one direction (say, from left to right), whereas on the lattice, there are many directions possible. Following Cressie (1993), page 405, let us denote the square integer lattice by $D = \{s = (u, v)^T : u \in \mathbb{Z}, v \in \mathbb{Z}$ i.e. s is an ordered pair indexing the positions on the lattice. If there are N

regions on the lattice they are indexed by $s_i, i = 1 \dots N$. The first approach, called simultaneous SAR, dates back to Whittle (1954). Let $\{Z(s) : s \in D\} \supset \{Z(s_i) : i = 1 \dots N\}$ is defined on the lattice D. Let also $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Lambda)$ be a n-dimensional joint Gaussian distribution with mean vector $\mathbf{0}$ and a diagonal covariance matrix i.e. $\Lambda = \sigma^2 I$ where I is the identity matrix of dimension $N \times N$. Also $W = \{w_{ij}\}$ is the $N \times N$ weight matrix as in (6.7) and the SAR model for $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))^T$ can be expressed in matrix form:

$$(I - W)(\mathbf{Z} - \boldsymbol{\mu}) = \boldsymbol{\epsilon}. \tag{6.8}$$

For this model to hold, it is assumed that $(I - W)^{-1}$ exists. Details can be found in Ripley (1981). A distinguishing feature from the time-domain UAR is that there is a dependence between the errors of the model and the lagged variables i.e. $cov(\boldsymbol{\epsilon}, \mathbf{Z^T}) = E(\boldsymbol{\epsilon}\mathbf{Z^T}) = \Lambda(I - W^T)^{-1}$ which is not diagonal, unlike the UAR case. Consequently, least-squares estimators of the parameters will not achieve consistency (Whittle (1954)).

Another possible specification of SAR restricts the possible dependence, so that the issue with the correlation of residuals and lagged regressors is mitigated. This is called the conditionally specified spatial Gaussian model. It includes a restriction on only *pair-wise* dependencies in the lattice and using Markov properties for the conditional density. We could think of it as paths on a graph, where every entry is conditionally independent of all others given its neighbours. Details can be found in Cressie (1993), page 407.

Similarly, the extension of the whole plethora of time series ARIMA models to spatial ARIMA is more complicated e.g. we could have a space-time moving average, a regression-type spatial ARMA i.e. if we have regressors in (6.8), then, say, $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$ can be non-trivial — to fit the mean of a spatial dataset correctly — exactly because of the multiple directions of the dependencies possible. For example, let our spatial correlation coefficient from equation (6.5) be a scalar ρ. Then equation (6.8) would simply become:

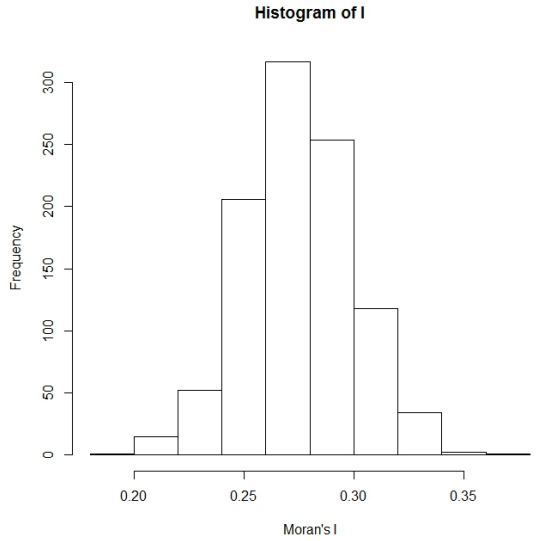

Figure 6.14: The empirical distribution of Moran's I for a SAR model with $\rho = 0.5$ for 1000 realizations from a 32×32 grid

$$(I - \rho W)(\mathbf{Z} - \boldsymbol{\mu}) = \boldsymbol{\epsilon}. \tag{6.9}$$

Thus, we can say the the dependence from the autocorrelation coefficient ρ is transferred to the generated data via the weights matrix W. Furthermore, this can become very complicated — imagine that ρ was not a scalar, but rather an $N \times 1$ vector , then this would correspond to spatially varying autocorrelation.

6.7.2 Relationship between Moran's I and SAR parameter ρ

As already explained in the previous section, the SAR model from (6.9) is different from the standard UAR model. For example, Fig. 6.14 shows that most of the time (for a nearest neighbour grid weighting scheme) I is half of ρ — the empirical mean of I is 0.27 for 10^3 realizations with $\rho = 0.5$. This can intuitively be explained by the mere fact, that in the spatial world, "time" goes in two directions rather than one as in univariate time series.

Simulation results			
Grid	Moran	HWWN	HWWN(BH)
16 by 16	0.059	0.01	0.03
32 by 32	0.046	0.02	0.04
64 by 64	0.044	0.03	0.05
128 by 128	0.047	0.03	0.05

Table 6.1: Statistical Size for $MVN \sim (\mathbf{0}, I)$ for Moran's test, our HWWN and HWWN(BH) - with using False Discovery Rate instead of Bonferroni.

6.7.3 Empirical size simulations

Here we are going to compare the 2D *HWWN* and the Moran test (using the function `moran.test` from `spdep` R package). For the Moran test, we also have to specify the weighting scheme. We chose to have 4 links for each observation on the grid ("rook" type weights) and the "torus" option for boundary conditions. The model is (6.8). We evaluated both tests' statistical sizes by simulating from the white noise model $\mathbf{Z} \sim MVN(\mathbf{0}, I)$ where I is the identity matrix as the covariance matrix. The empirical sizes (with 10^3 replications) of our test from section 6.4 and Moran's test can be found in Table 6.1. We can conclude that Moran's test matches statistical size, whereas our 2D HWWN is a bit conservative. We also used another version of our test, where in the last step in the procedure from section 6.4 we used false discovery rate (FDR) from Benjamini and Hochberg (1995b), rather than Bonferroni — in that case, the empirical size is closer to the theoretical size.

As a conclusion, our test is a bit conservative, whereas Moran matches statistical size. However, this is a first version of our test and it might be further tuned (we might take only the finest scale coefficients for example). We tried the same simulation, for Moran, using "queen" type contiguity which translates to 8 links per observation on the grid and its size largely remained the same. For illustration,Fig. 6.15 shows that the distribution of the raw p-values of all the coefficients tested (16380) from 2D *HWWN* seems to be uniform.

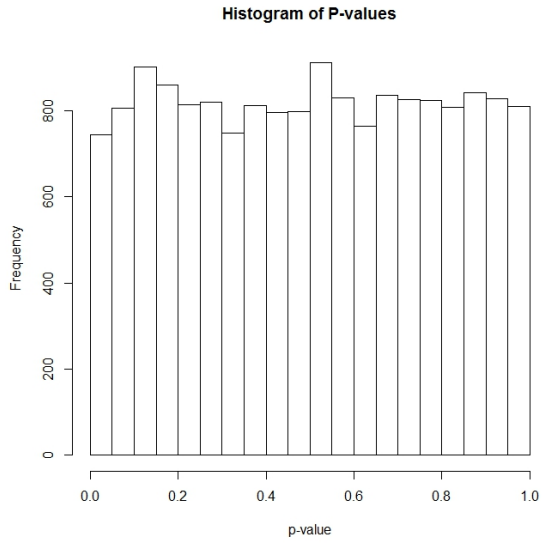

Figure 6.15: Histogram of ordinary p-values from 2D *HWWN* test of a 128×128 white noise data — for all coefficients at all scales

Simulation results		
Grid	Moran	HWWN
8 by 8	0.06	0.01
16 by 16	0.06	0.02
32 by 32	0.05	0.03
64 by 64	0.06	0.03
128 by 128	0.05	0.04

Table 6.2: Statistical Size for $U \sim (0, 1)$

Data	Simulation results	
Grid	Moran	HWWN
8 by 8	0.30	0.42
16 by 16	0.88	0.76
32 by 32	1.00	1.00

Table 6.3: Statistical Power for Gaussian UAR with $\rho = 0.3$

Data	Simulation results	
Grid	Moran	HWWN
8 by 8	0.60	0.82
16 by 16	1.00	1.00

Table 6.4: Statistical Power for Gaussian UAR with $\rho = 0.5$

Another layer of validation of the performance of the 2D *HWWN* test for spatial statistics is if we use uniform distribution. Again, under the null hypothesis, we generated n realizations of $U(0, 1)$ distributions for the columns of the matrix $\mathbb{X}_{ts}$. The empirical size of 2D *HWWN* for uniform noise on the interval $(0, 1)$ is 4.1%. For different number of observations, a range of simulations appears in Table 6.2. So, we can conclude that Moran matches theoretical size, but sometimes exceeds, whereas our 2D *HWWN* is slightly conservative.

6.7.4 Empirical power simulations

For statistical power, firstly we used series of independent univariate AR(1)(of variable length n and different values of the parameter ρ_1) process for all columns of our input $n \times n$ matrix and both tests reach 100% power when the magnitide of the parameter and the number of observations are increasing. A range of powers for different AR1 parameter values is available in tables 6.3, 6.4 and 6.5. Similar features for univariate white noise

Data	Simulation results	
Grid	Moran	HWWN
8 by 8	0.82	0.98
16 by 16	1.00	1.00

Table 6.5: Statistical Power for Gaussian UAR with $\rho = 0.7$

Data	Simulation results	
Grid	Moran	HWWN
16 by 16	0.63	0.01
32 by 32	0.92	0.03
64 by 64	1.00	0.03

Table 6.6: Statistical Power for Gaussian SAR with $\rho = 0.2$

Data	Simulation results	
Grid	Moran	HWWN
16 by 16	0.64	0.01
32 by 32	1.00	0.97
64 by 64	1.00	1.00

Table 6.7: Statistical Power for Gaussian SAR with $\rho = 0.5$

tests were explored in the simulations in chapter 3. We will note however that this does not corresponds to (6.9), since we are inducing dependence only in one direction, rather than two. Thus it is a more simplistic scenario. However, in spatial regression residual analysis for example, there would be certain, say small, regions which could exhibit auto-correlation and are of interest. Thus, as a future work, it might be interesting to simulate such processes.

Next, we are going to do some comparisons for a genuine SAR example. We use (6.9) with different scenarios for both the grid size and the magnitude of the SAR parameter ρ — 0.2, 0.5 and 0.8. We can conclude that Moran's test is very good when the grid is of small size, whereas our test performs very poorly. When the SAR parameter is of low magnitude Moran performs reasonably well, whereas HWWN does not do well. Those results also confirm the finding from Cliff and Ord (1972) that Moran's test performs very well when the parameter is of small magnitude. This assertion has also been explored in Robinson and Rossi (2014).

Looking at Tables 6.7 and 6.8, we notice a jump in the empirical power from 16×16 to 32×32 grid. Unofortunately, we cannot test 24×24 grid size due to the dyadic wavelets we use. However, we did additional simulations with $\rho = 0.4$ for a 32×32 grid and the power of our test is 70%.

Data	Simulation results	
Grid	Moran	HWWN
16 by 16	0.60	0.01
32 by 32	1.00	0.97
64 by 64	1.00	1.00

Table 6.8: Statistical Power for Gaussian SAR with $\rho = 0.8$

Data	Simulation results	
Grid	Moran	HWWN
16 by 16	1.00	0.06
32 by 32	1.00	0.33
64 by 64	1.00	1.00

Table 6.9: Statistical Power for Gaussian SAR with $\rho = -0.3$

We did also simulations with negative values of the SAR parameter — in Tables 6.9 and 6.10 — the results are good and show that it is easier for both Moran and our test to reject the null hypothesis of white noise when the parameter is negative.

6.7.5 Refining the test as univariate d_{22}

We have better results when using the $2D$ analogue of the univariate d_{22}. To add.

6.7.6 Homogenizing and contaminating of white noise

In this section we will do a simple experiment with grid data. We will start with a matrix of 128×128 matrix of Gaussian data with mean three and standard deviation of two and will try to homogenize and contaminate it. The mean is not zero on purpose and standard deviation is different from one for better visualization.

Homogenizing will be the following operation:

Data	Simulation results	
Grid	Moran	HWWN
16 by 16	1.00	0.29
32 by 32	1.00	0.99
64 by 64	1.00	0.99

Table 6.10: Statistical Power for Gaussian SAR with $\rho = -0.5$

1. Starting from 128×128 white noise (Fig. 6.16), take the upper and lower diagonals and assign them the mean value of the main diagonal.

2. Then repeat this for off-main-diagonals 2 to 7, assigning the mean of the previous diagonal i.e. for the second diagonal, assign the mean of the first off-main diagonal and so on to the seventh diagonal.

Then we will do 10^3 simulations to see if our test and Moran's test still not reject the null hypothesis. We are trying to assess how the 2D HWWN and Moran's test sizes would respond to a systematic statistical effect without adding new data.

For illustration, Fig. 6.16 shows the image of the generated white noise of dimensions 128×128. Then, on Fig. 6.17 we can see what happens to the image after the described *homogenizing* operation. We evaluated the empirical size — Moran's size is 5.6% and for our 2D HWWN test is 2.8% which corresponds to the results from other simulations in the previous section. Similarly, the empirical size for our test, with using FDR instead of Bonferroni, is 6.7%. This conforms to the results from the previous section and we will expect the using FDR would increase the empirical power of the test, at the expense of light oversizing.

Contaminating will be the following operation:

1. Starting from 128×128 white noise (Fig. 6.16), take the upper and lower diagonals and replace their entries with a realization UAR with parameter $\rho = 0.5$.

2. Then repeat this for off-main-diagonals 2 and 3, replacing them with UAR realizations with parameters of 0.4 and 0.3 respectively.

Then we will do 10^3 simulations to estimate the empirical power of 2D HWWN and Moran's test . We are trying to assess if both tests would detect would respond to a systematic statistical effect when a fraction of non-white noise datapoints i.e., $2(127 + 126 + 125) = 756$ out of $16384(128 \times 128)$, are inserted throughout regions of the lattice.

Fig. 6.18 shows the red "motorway" with the alternating colours line in the middle, since the main diagonal was kept unaltered. The empirical power of Moran is 100% and ours is 85%. However, if we use the FDR method from Benjamini and Hochberg (1995b), then our power is 95%. Those results conform to our previous findings regarding the power of our 2D HWWN test.

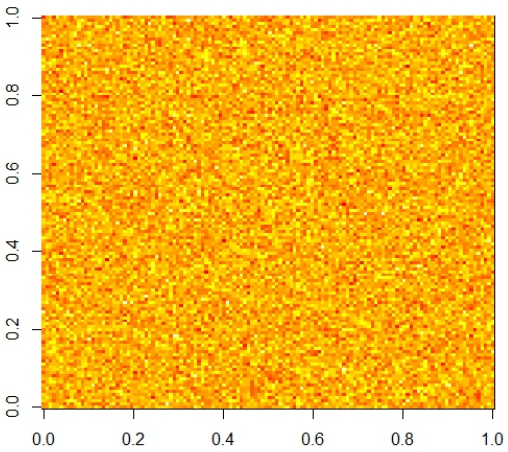

Figure 6.16: Image of 128×128 observations of Gaussian white noise with mean of 3 and variance 4

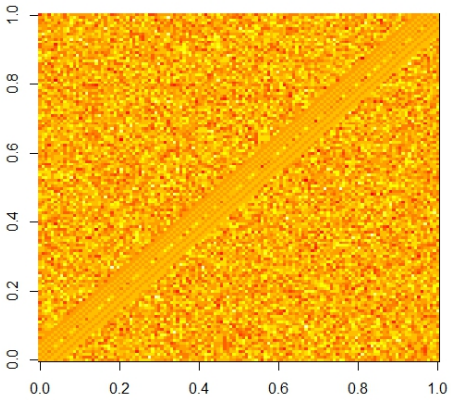

Figure 6.17:
Image of 128×128 Gaussian white noise, after the *homogenizing* operation

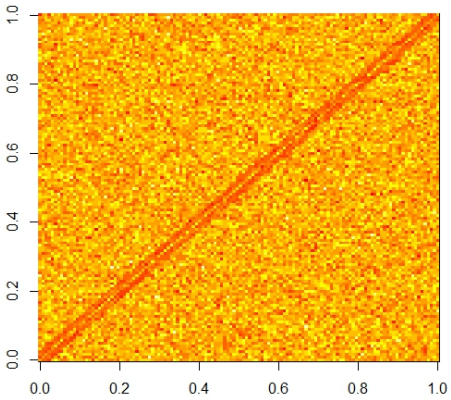

Figure 6.18:
Image of 128×128 Gaussian white noise, after the *contaminating* operation

6.8 Conclusion and Further Work

In this chapter we developed a two-dimensional wavelet test for white noise using Haar wavelets. We also analysed the performance of the classical spatial autocorrelation test of Moran (1950) and our HWWN test against spatial autoregressive alternative hypotheses and custom scenarios. The results are encouraging since our test performs well for empirical size and reasonably for empirical power for a range of magnitudes of the spatial autoregressive model (6.9) parameter ρ. As a first version, our two-dimensional test has some deficiencies, especially low power when the size of the grid is small and when the SAR parameter is of low magnitude. However, those flaws might be expected given the preliminary form of the test.

For example, due to our replication of the positive frequencies of the two-dimensional periodogram, we are inducing more tests than is needed. Undoubtfully, this operation has affected the statistical power of the test. This could be alleviated by selecting only the wavelet coefficients that correspond to the "first half" of the replicated periodogram since that would reduce the number of multiple comparisons the test is evaluating. Another option is to correct the nominal size of the test.

Another direction for research might be the orthogonal two-dimensional Haar wavelet transform itself. For example, since spatial autocorrelation is a phenomenon that has multiple facets, the use of the non-decimated wavelet transform might yield more powerful results. Moreover, there exist more complicated two-dimensional wavelet transforms such as wavelet packets which could provide further insight into the two-dimensional white noise testing problem.

Eventually, we believe that wavelet-based spectral tests have prominent future in spatial statistics, and this chapter represents one attempt of this endeavour.

Chapter 7

ARH Order Verification Methodology

7.1 ARH(1) Order Verification

7.1.1 Background

A key point for ARH(1) processes is that the autoregressive structure is propagated through all the coordinates of the process with respect to the Karhunen-Loeve basis. The main result from Bosq (2000), concerning the propagating structure is Theorem 3.6, which we reproduce here.

Definition 15. *Symmetric Operator in Hilbert Space H*

An operator l, defined on a Hilbert space H, is symmetric if

$$< l(x), y >=< x, l(y) >, \qquad x, y \in H. \tag{7.1}$$

Note: A linear operator on an infinite-dimensional space can be expressed as an infinite matrix. However, with our discretized functional data, the matrices will always be finite thus the discretized lag one autocorrelation operator will actually be a lag one autocorrelation matrix of dimensions $p \times p$, where p is the length of the grid on which our realizations are recorded. Also, symmetric in the sense of Definition 15 will just mean

that a finite square matrix that is symmetric.

Suppose we have the ARH(1) model: $(\mathcal{X}_n) = \rho(\mathcal{X}_{n-1}) + \varepsilon_n$ as defined in Def. 14, where ε_n is defined as in Def. 13. Let $\{e_j\}$ be the Karhunen-Loeve eigenbasis of the functional autoregressive operator ρ i.e. $\rho(e_j) = \alpha_j e_j, j \geq 1$ and $\lim_{j\to\infty} |\alpha_j| = 0$ is a decreasing sequence of numbers.

Theorem 4. *(Bosq(2000), Theorem 3.6)*

If $\boldsymbol{\rho} = \sum_{j=1}^{\infty} \alpha_j e_j \otimes e_j$ *is a symmetric compact operator over H, then* $(\mathcal{X}_n)$ *is a zero-mean ARH(1) associated with* (ρ, ε) *if and only if*
$(< \mathcal{X}_n, e_j >)$ *is an (eventually degenerate) AR(1) associated with* $(\alpha_j, < \varepsilon_n, e_j >)$ *where* $<,>$ *denotes dot product in the Hilbert space* $H = L^2$ *space.*

A note on the implementation of Theorem 4

The ARH(1) process is defined through its operator i.e. $\mathcal{X}_n = \rho(\mathcal{X}_{n-1}) + \varepsilon_n$. The parameter's ρ Euclidian norm determines the process i.e. if $\|\rho\|_{L_2} = 0$, then the process is functional white noise. The condition for validity of the ARH(1) process, see (3.14) is $\|\rho^{j_0}\|_{\mathcal{S}} < 1$ where j_0 is an integer and $\mathcal{S}$ denotes the norm in the space of Hilbert-Schmidt operators as in Def.11 from the literature review. However, when we work with discretized functional data $\mathbb{X}_{n\times p}$ the operator ρ is a $p \times p$ matrix thus, the Hilberd-Schmidt norm is the Frobenius norm which is the same as the L_2 norm. That is why, the α_j in Theorem 4 are a decreasing sequence and are required to be in $(-1, 1)$ (so that the autoregressive corrdinates (i.e. the principal components), the η_j from equation (7.3), in the eigenbasis are valid stationary AR1 processes), see section 7.1.2, since $\alpha_j, \quad j = 1, \ldots, p$ are the univariate autoregressive parameters that are the result of the projection of ρ on the eigenbasis.

7.1.2 Our methodology for verification of ARH(1)

Usually, when an eigendecomposition of an ARH process is obtained, its validity is judged by the amount of variance explained through the eigenvalues and various cross-validation (CV) scores. There are different ways to do that as outlined in Besse et al. (2000) and Bosq (2000). Another important question, that we adress in this chapter, is how do we conclude that a particular process is of the ARH(1) kind similar to order verification in the scalar and multivariate cases.

Theorem 4 is a theoretical and idealized case, because the lag one autocorrelation operator $\boldsymbol{\rho}$ of such a process could only be symmetric in the sense of Def.15 in a limiting case. Nevertheless, Theorem 4 might have an interesting practical application within the context of discretized functional time series. The rationale is that, in theory, when both the number of observations n and the number of gridpoints p goes to infinity or, in practice, it is sufficiently large, then we will be close to the asymptotic case of Theorem 4. Then we can use the symmetrized version of our autocorrelation lag one matrix in a particular way. Here we are considering verification of order for ARH(1) process as defined in (3.14).

Suppose we have $n \times p$ discretized functional data denoted by $\mathbb{X}_{np}$ that come from an ARH(1) process. Consider the lag-1 autocorrelation matrix $\boldsymbol{R_1}$ of those data and then form its symmetrized version which will be denoted $\boldsymbol{R_1^{sym}}$:

$$\boldsymbol{R_1^{sym}} = \frac{\boldsymbol{R_1} + \boldsymbol{R_1^T}}{2} \tag{7.2}$$

Now, we can form the principal components which are defined as the projections of the data onto the eigenbasis of the matrix $\boldsymbol{R_1^{sym}}$ (which is of dimension $p \times p$). Let the eigenvectors of $\boldsymbol{R_1^{sym}}$ be denoted by e_j for $j = 1 \dots p$, then the principal components are:

$$\eta_j = \mathbb{X}_{np} \times e_j. \tag{7.3}$$

Here $\times$ denotes matrix multiplication. Now we will apply Theorem 4. As the data are of dimension $n \times p$ and the eigenvectors e_j are of dimensions $p \times 1$, then the principal components are of dimensions $n \times 1$ i.e. they are just projections of the original $\mathbb{X}_{np}$ data on the eigenbasis. Theorem 4 implies that each principal component η_j for $j = 1 \dots p$ follows an AR(1) process in the form $\eta_{j_{t+1}} = \alpha_j \eta_{j_t} + \varepsilon_{j_t}$ where α_j is a number $\in (-1, 1)$, $t = 1, \dots, n-1$ and ε_t is Gaussian white noise. Based on this, we propose the following statistical procedure for verification of ARH(1) in the next section.

7.1.3 Algorithm for verification of ARH(1) — *VERARH*

1. Form the $\boldsymbol{R_1^{sym}}$ matrix and its spectral decomposition.

2. Perform a principal component analysis with respect to $\boldsymbol{R_1^{sym}}$ and project the data on each of its eigenvectors which will give p principal components — the η_j.

3. Assume the model $\eta_{j_{t+1}} = \alpha_j \eta_{j_t} + \varepsilon_{j_t}$ for each principal component then test:
   H_0: The series η_{j_t} is white noise versus
   H_a: The series has an autoregressive structure, up to lag $\log n$.
   Test this using Ljung-Box (L-B) with lag $\log n$ and size $s/p\%$ that accounts for multiple testing as per Bonferonni correction method.

4. Record the number, k, of statistically significant principal components from the previous step.

5. Take the k components from the former step and fit scalar AR(1) process to each of them.

6. Test the residual time series from each principal component and do the same test, using Ljung-Box up to lag $\log n$, correcting the degrees of freedom. In case the selected k residuals of principal components are determined white noise, then the process is ARH(1).

The principal components correspond to the ordered eigenvalues, thus representing variance in the functional data. A suggested threshold in Kokoszka and Reimherr (2013) is to select the principal components that account for at least 90% *of the variance which we apply here.*

Note that in step 3 we are using higher lag, in case the process is of higher order. In step 4 we are testing all the principal components that have rejected the null hypothesis, however, we can test only the ones that cumulatively account for 90% of the variance.

Verification of ARH(p)

In chapter 5 of Bosq (2000), there are limiting theorems similar to Theorem 4, but for higher orders of the ARH process. Therefore, we can design a more general test for order ARH(p) via successive AR fitting to the principal components (equation 7.3), for which an example will be shown section 7.3.1.

7.2 Simulation study for ARH(1)

7.2.1 Setup of the Simulation Routine for ARH(1)

So, we work with discretized data and the operator ρ will be approximated by a matrix with dimensions $p \times p$, where p is the number of discretization points. The simulation routine `simul.far.wiener` computes the operator in the Karhunen-Loeve basis which, in the matrix case, is the eigenbasis. This means that the η_j are the projections of the data over the eigenbasis, using the eigenvectors of the covariance/correlation matrix. The linear operator — ρ — in this basis, i.e. a $p \times p$ matrix, admits bloc structure of the form:

$$\rho = \begin{pmatrix} d.rho & 0 \\ 0 & eps.rho \end{pmatrix} \tag{7.4}$$

Here *d.rho* is provided in the call of the function and defines the α_j from Theorem 4

i.e. say we have p principal components which is also the number of discretization points, then we decide that the scalar value for the autoregressive parameters for the first k of them is a number, thus our *d.rho* bloc will be just `rep(value,k)`. This corresponds directly to having k number of scalar AR(1) process in the eigenbasis admitting the form $\eta_{j_{t+1}} = \alpha_j \eta_{j_t} + \varepsilon_{j_t}$ for $j = 1 \dots k$ and $t = 1 \dots n-1$. The *eps.rho* bloc is providing the dependence in the rest $p-k$ principal components via the following perturbation scheme:

$$rho.eps = (\epsilon_{k+1}, \epsilon_{k+2}, \dots, \epsilon_{2p}) \qquad where \tag{7.5}$$

$$\epsilon_i = \frac{perturbation}{i^2} + \frac{1 - perturbation}{e^i} \tag{7.6}$$

For $i = k+1 \dots 2p$ and perturbation is a perturbation coefficient, by default equals 0.05 which means that the expression will drop fast from the predefined values. The expansion in the eigenbasis goes up to $2p$ and to get the discretized functional data, the routine then projects the $n \times 2p$ data from the eigenbasis to $n \times p$ data in the canonical basis. For more details please see the help of the `far` package.

Simulation scenario setup

1. Generate 1000 realizations, each consisting of $n = 500$ curves and $p = 100$ discretization points.

2. Use 2 parameter values (for *d.rho* i.e. α_j) — 0.8 and 0.5 for half of the principal components $\eta_{j_t}, \quad t = 1, \dots, 50$ i.e. in the call of the simulation routine we will use `rep(0.8, 50)` and `rep(0.5, 50)` for *d.rho* respectively.

3. Apply *VERARH* from section 7.1.3 — for step 6 with the Ljung-Box test, up to lag 6, on each of the residuals from AR(1) fitting to the 100 principal components for every realization and the tested null hypothesis will be $d.rho[i] = \alpha_i = 0$ and the alternative $d.rho[i] = \alpha_i \neq 0$ for $i = 1, \dots, 100$. The nominal sizes are 5×10^{-7}

and 1×10^{-7} and they also account for multiple comparisons by Bonferonni correction scheme(10^{-3} for realizations $\times 10^{-2}$ for principal components $\times 5(1).10^{-2}$ for nominal size).

4. For each of the sizes its empirical power will be calculated.

7.2.2 Results from applying VERARH for the verification of ARH(1)

We can see the results in table 7.1 are very good and confirm our assertion that our *VERARH* procedure can be used for verification of ARH(1) for functional time series. By the empirical size for the performed tests, we can conclude that the test is very conservative and reaches 100% power.

α_j	Empirical size	Power
0.8	1×10^{-7}	1.00
0.8	5×10^{-7}	1.00
0.5	1×10^{-7}	1.00
0.5	5×10^{-7}	1.00

Nominal size ($\alpha_j = 0$)	Empirical size
0.05	0.0012
0.01	0.0004

Table 7.1: Empirical power (left) and size (right), testing 100 principal components for 1000 realizations of ARH(1) with *VERARH*

It is also illuminating to look at the empirical distributions of the autocorrelation coeffient(α_j) of each principal component. Figures 7.1 and 7.2 show the α_i and their respective p-values. We see that the first 50 principal components have their autoregressive paramaters (autocorrelation coefficients) centered around their respective predefined values, 0.8 and 0.5. Then from principal component 50 onwards the perturbation scheme comes into effect. This is way more visible in the right pictures of Figures 7.1 and 7.2 where p-values for principal components 50 onwards become larger and dispersed.

7.2.3 Conclusions from the simulation Results

The proposed *VERARH* methodology, from section 7.1.3, for verification of ARH(1) process is feasible and reliable as shown and validated from simulated data. Usually, when doing principal components analysis of functional data, we get most of the variance (80 to 90 %) concentrated in the first 3 to 5 eigenvalues — Hörmann and Kokoszka (2012). This means that doing 5 tests would be enough in practice to determine whether or not a functional dataset follows an ARH(1) process.

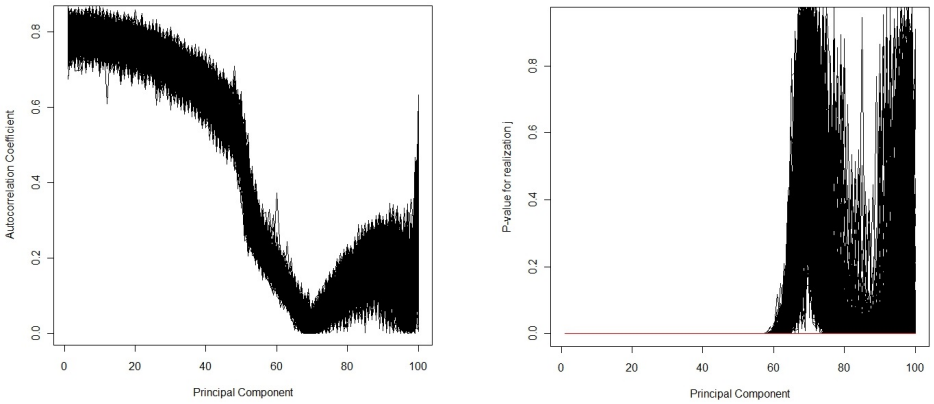

Figure 7.1: left: distribution of the value of the autocorrelation coefficient, defined as $\alpha_j = 0.8$ for the first 50 pricinpal components and right: its p-value, for all 1000 realizations with $n = 500$ curves and p=100 discretizations points/principal components

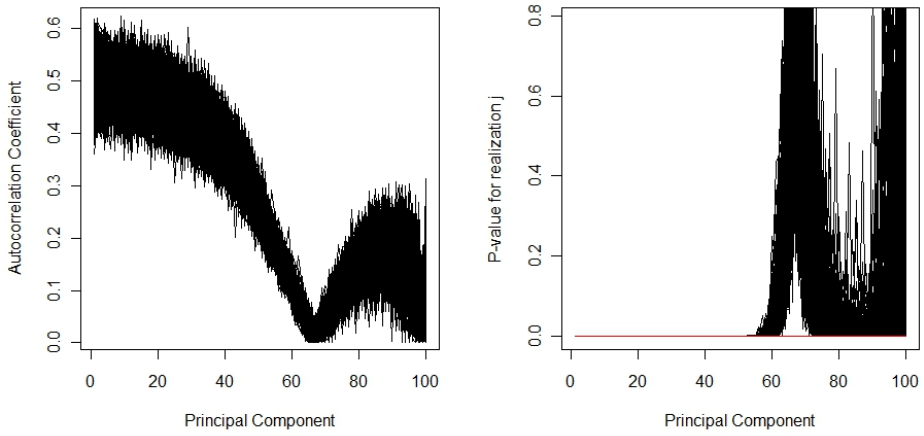

Figure 7.2: left: distribution of the value of the autocorrelation coefficient, defined as $\alpha_j = 0.5$ for the first 50 pricinpal components and right: its p-value, for all 1000 realizations with $n = 500$ curves and p=100 discretizations points/principal components

7.3 Simulation Study for ARH(2)

This section considers modifying testing procedure *VERARH* by using the d00 single coefficient Haar wavelet test from chapter 3 instead of Ljung-Box. The modification of the *VERARH* is provided in section 7.3.3Furtheremore, we introduce slight changes in the multistaging procedure in order to be ables to verify higher orders than one for ARH. We will compare our new procedure to the one from Kokoszka and Reimherr (2013) by simulation. Furthermore, Kokoszka and Reimherr (2013) procedure is also multistage in the sense that it tests sequentially ARH(0) vs ARH(1), then ARH(1) vs ARH(2) and so on. We will also show an empirical size calculation when doing the ARH(2) simulation study in this section.

In order to be able to test for ARH(p), the modifications of our algorithm from section (7.1.3) consist of the following: *Extended VERARH*

- In step 5 fit an AR(2) process to each of the k statistically significant principal components from step 4
- In step 6 consider the residuals from the AR(2) for making the decision.

7.3.1 Setup of the Simulation for ARH(2)

We are not aware of any R packages that have a routine for simulating ARH(2) process. That is why here we used the procedure mentioned in Kokoszka and Reimherr (2013). We express our gratitude to the authors for providing it.

The ARH(2) is defined in the following way:

Definition 16. *ARH(2) process*

$$\mathcal{X}_n = \rho_1(\mathcal{X}_{n-1}) + \rho_2(\mathcal{X}_{n-2}) + \varepsilon_n \qquad n \in \mathbb{N} \tag{7.7}$$

size	Power
1×10^{-5}	100%
5×10^{-5}	100%

Table 7.2: Results from Ljung-Box test up to lag 6, testing the residuals of AR(2) fits to the first 5 principal components for 1000 realizations of ARH(2)

where ρ_1 and ρ_2 is an infinite-dimensional linear autocorrelation operators and ε_n is a Brownian Bridge.

Note on the operators ρ_1 and ρ_2

In this simulation they are again discretized as $p \times p$ matrices and the crucial quantity is the norm. As explained in the appendix **??** we are considering the L_2 norm and this is the quantity that governs the ARH process, be it order one (as in Definition (3.14)) or two as just defined in Definition(7.7). However, for ARH(2), we have an additional condition for stationarity, similar to scalar AR processes, that the sum of the norms of the two operators must be less than one, for the ARH(2) model to exist.

The simulation will be performed as follows:

1. Generate 1000 realizations, each consisting of $n = 500$ curves and $p = 100$ discretization points. We use a burn-in of 200 curves.

2. The L_2 norm of the operator ρ_1 is 0.6 and for ρ_2 it is 0.3

3. The extended VERARH will be performed with sizes — $0.05/5000$ and $0.01/5000$, as per Bonferonni correction scheme for the empirical power to be calculated.

7.3.2 Results from applying extended VERARH for the verification of ARH(2)

Note: In Fig.7.3 for all the 1000 realizations we always had that over 90% of the variance is in the first 5 principal components that is why the nominal sizes are such. The empirical size is the same as in the ARH(1) simulation, so we will not consider it here.

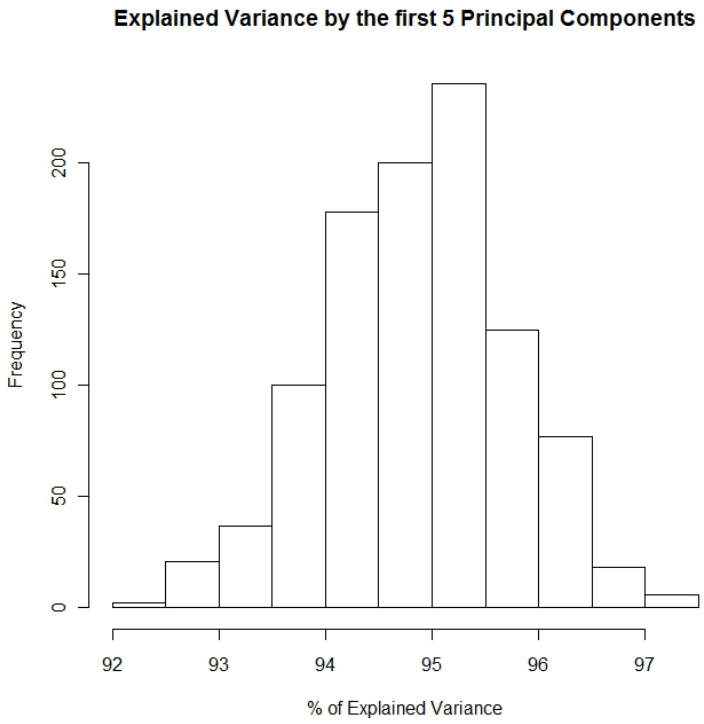

Figure 7.3: Percentage of explained variance by the first 5 principal components for 1000 realizations of ARH(2)

7.3.3 The upgraded VERARH procedure, using d00 wavelet test

Algorithm for verification of ARH(m) — VERARHm

1. Form the regular correlation matrix and its spectral decomposition.

2. Perform a principal component analysis i.e. project the data on each of its eigenvectors which will give p principal components — the η_j.

3. Assume the model $\eta_{j_{t+1}} = \alpha_j \eta_{j_t} + \varepsilon_{j_t}$ for each principal component with H_0: The series is white noise versus

   H_a: The series is not white noise.

   Test this using d00 with size $s/p\%$ that accounts for multiple testing as per Bonferonni correction method.

4. Record the number, k, of statistically significant principal components from the

previous step.

5. Take the k components from the former step and fit scalar AR(1) process to each of them.

6. Test the residual time series from each principal component and do the same test, using d00

The principal components correspond to the ordered eigenvalues, thus representing variance in the functional data. If the residuals from the AR(1) fits of the principal components, accounting for most of the variance — 90%, do not reject the null hypothesis from d00 test, then decide that the process is ARH(1).

If, in step 6, we reject the null hypothesis, the we conclude we have ARH order greater than one. To find the correct order we have to re-iterate through the procedure:

- In step 5 fit an AR(2) process to each of the k statistically significant principal components from step 4.

- In step 6 consider the residuals from the AR(2) for making the decision.

This procedure can be adapted to any reasonable order. Now, let us see it in practice. We will do a comparison with Kokoszka and Reimherr (2013) for the ARH(2) example from the paper, with norms of 0.5 and 0.3 for ρ_1 and ρ_2 respectively which corresponds to c_1 and c_2 from model(18) in their paper and their Table 2. We generated 200 curves with 100 gridpoints, a burn-in of 200 curves and 10^3 replications. For statistical size, our procedure is conservative — for nominal 5% level, the size we get is 1.3% and theirs is 5.9%. The statistical power, when comparing ARH(0) versus ARH(1), is 100% for both ours and Kokoszka and Reimherr (2013). The power, when comparing ARH(1) versus ARH(2), we get is 93.4% while they get 96.1%. Furthermore, we tried to reiterate and check with the upgraded VERARH for ARH(2) vs ARH(3), and (as this becomes a statistical size calculation) the size was 0%.

Conclusion from ARH(2) simulation study

So, based on this simulation experiment, and comparison with the method of Kokoszka and Reimherr (2013), we could conclude that the suggested extended version of the *VE-RARH* algorithm, defined in section 7.3.3, using d00 test, is a reliable tool for verification of the order of ARH processes.

7.3.4 A Real Functional Time Series Example

Description of the Data and the Problem

The Elelctricity de France(EDF) data are observations of the electricity load of the EDF network in France. They start on 1 Sep 2002 until 31 Aug 2009, which equates to 2557 days. The data are sampled every half-hour i.e. there are 48 observations for every day. The functional representation of the data is a curve for each day, comprised of 48 points on a grid. So, we will have a matrix $\mathbb{X}_{np}$ of $n = 2557$ rows (these represent days) and $p = 48$ columns (these representing the half-hour intraday sampling). We will use the term daypoints/curves to refer to the rows of the matrix and the term gridpoints to refer to its columns.

These type of data are very popular in the functional time series literature. For instance, in Cho et al. (2013), they have been modelled with Generalized Additive Models and Dimension Reduction Techniques. Whereas, in Antoniadis et al. (2010), they have been modelled with ARH processes, however, after complicated curve clustering procedures. This is usually necessary as curves from real processes usually contain plenty of different modes of variation and nonstationarity. So, the clustering is needed to come with similar clusters in which one model could be fit for each. We will try to analyze the data as a whole, without incorporating other informations. This will allows us to see whether our procedure is adaptive enough to verify higher orders of ARH processes.

Figure 7.4 shows the daily data for EDF from 2002 till 2009. Figure 7.5 shows all

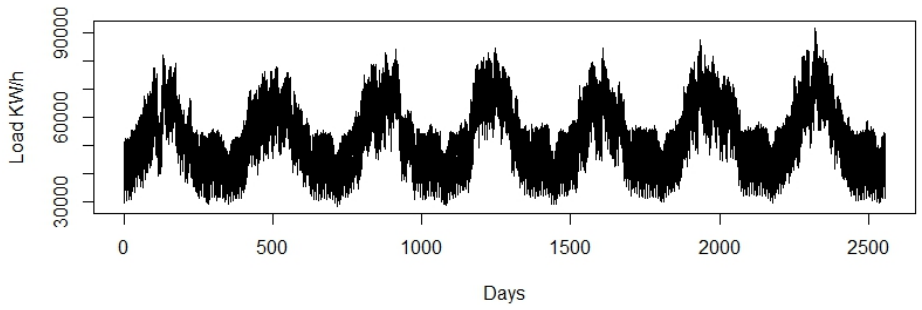

Figure 7.4: EDF daily electricity load time series from September 2002 till August 2009

the curves for the last one year of data. They look nice and smooth as curves, however, they are difficult directly to model with ARH(1) or ARH(2) because they possess strong correlation structure over intervals grater than lag 1. Moreover, as their vertical alignment suggests, they contain a lot of trends and increasing variance due to the seasonal character of the electricity consumption over the year, including intra-weekly, weekly, monthly, quarterly and year-to-year variation. That is why, for analysis and forecasting, we will use transformed data, using square root and then taking the lag 7 difference of order 1, both being done in this order and columnwise with respect to the multivariate $n \times p$ data matrix. The square root was chosen instead of the natural log, because it resulted in less correlations in lags other than 1 and 7. Lag 7 was particularly chosen as the most important variance driver, present in all seasons. For illustration Fig. 7.6 shows the autocorrelation and partial autocorrelation for 1st, 24th and 36th gridpoint representing midnight, noon and evening (6 pm) respectively.

Results from Verification Test on the Transformed Data

Looking closely at Fig. 7.6 we may notice that there is inherent periodic behaviour. Notably, lag 1 PACF has longer effect in periods of 7 lags, slightly decreasing. This might be a hint for monthly variation. On the other hand, there is a negative PACF at lag 7, repeating itself againg in periods of 7 lags and diminishing at the fourth instance. This is again in support of a monthly variation. It might be possible though that there are other

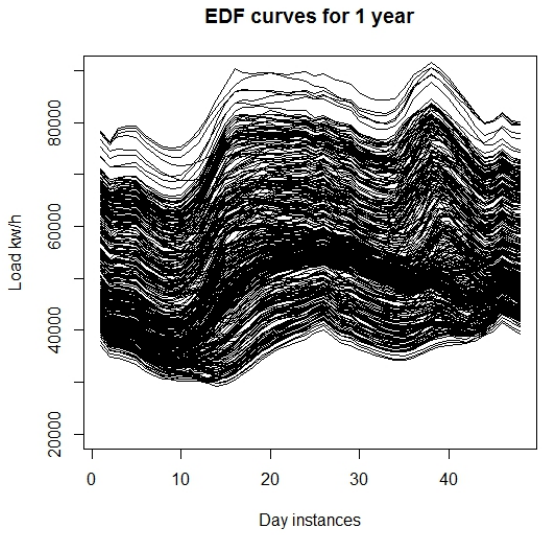

Figure 7.5: All the curves for the last 1 year of data i.e. Sep 2008 - Aug 2009

Item	Quantity	Share of Variance
Number of significant PCs	35	99.91%
Number of not significant PCs	13	0.09%
First 5 PCs by order of magnitude	5	99.57%

Table 7.3: Results from standard PCA and the method for the transformed EDF data.

transformations of the data which could help us get rid of this regular variation. However, we proceed as is, and this gives us a hint that probably ARH(1) would not suffice. We tried fitting successively AR(1) to AR(7) to the 48 principal components, however their residuals stil had structure. That is why, in step 5 of VERARH, we used AIC criterion in the *ar* procedure in R. It fitted AR models varying from 28 to 34 lags, which confirmed the monthly variation suggestion. In table 7.3.4 we can notice that the first 5 principal components are accounting for most of the variance. We could try fitting one single model to all 5 of them and then check if their residuals confirm the white noise hypothesis. We did fit an *AR(*31*)* to all of them and our residuals were not showing evidence of autoregressive structure anymore with the following p-values: $0.94, 0.92, 0.99, 0.98, 0.95$ for each of the 5 principal components respectively. So this agrees with the aforementioned references which show that these data are highly heterogeneous and require either higher order models or other techniques to be used in order to model them succesfully.

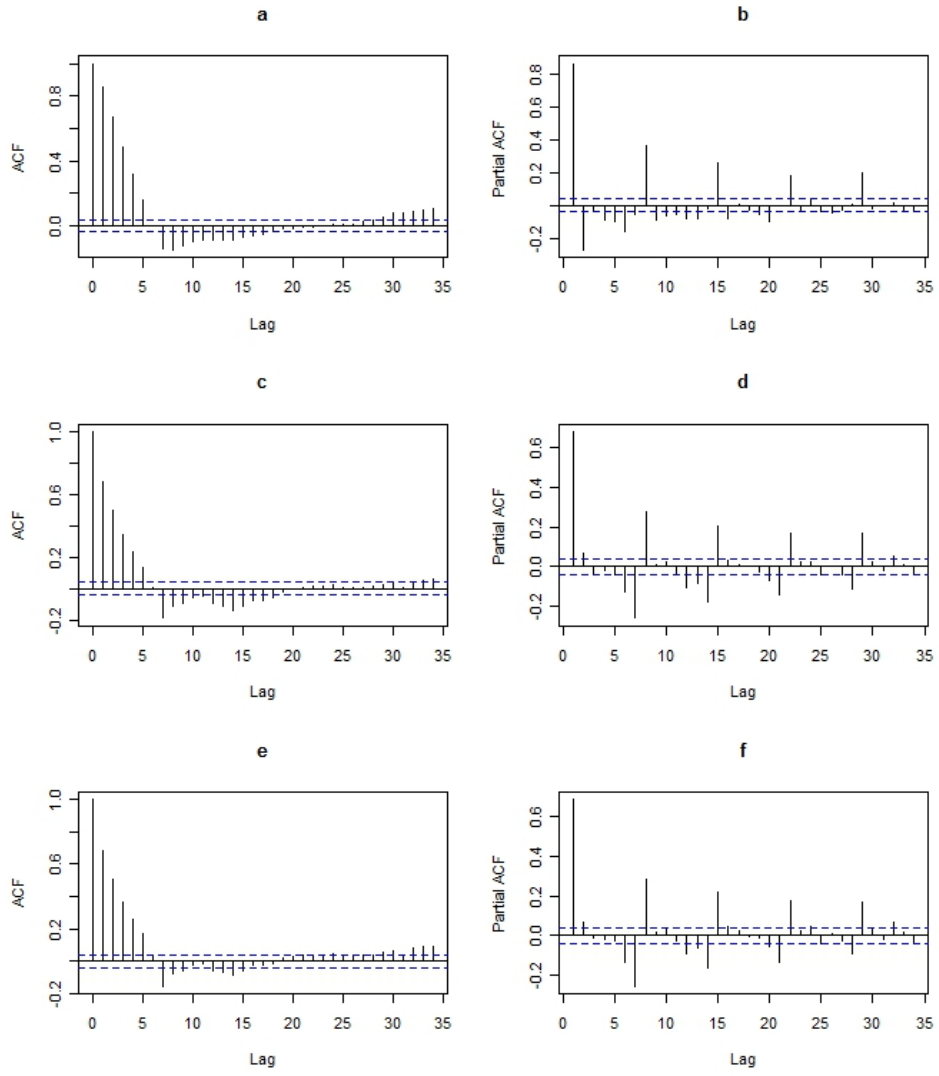

Figure 7.6: Autocorrelations for transformed daily data taken at the: a - 1st gridpoint(00:30 hrs), c - 24th gridpoint(12:00 hrs), e - 36th gridpoint(18:00 hrs)
Partial autocorrelations for transformed daily data taken at the: b - 1st gridpoint(00:30 hrs), d - 24th gridpoint(12:00 hrs), f - 36th gridpoint(18:00 hrs)

7.4 Simulation study with GCV and *VERARH* for ARH(1) prediction

7.4.1 Forecasting of ARH processes using the *VERARH* method

With the same arguments as in section 7.1.3, it is possible to select the number of principal components in the eigendecompositions of the regular and lag one covariance kernels. Then we could construct predictions using the principal components that conform to some rule.

When one is interested in forecasting functional time series, the most crucial question is the number of eigenvectors to use for projecting the original data as explored in (Bosq, 2000; Ferraty and Vieu, 2006; Besse et al., 2000; Damon and Guillas, 2005). There are at least two issues with this proposal. One is that there might be some measurement error in the data, as they are discretized functions. The second one is that, if only few principal directions of the variance (i.e. eigenvectors accounting for at least 90% of the variance) are retained, then the predictions are too smooth and very far from reality Hörmann and Kokoszka (2012). Those two reasons lead to the assertion that when using too few principal components, the predictions are too smoothed and do not catch the dynamics of the function well, but when using too many principal components, then we will be putting too much noise and our predictions might be incorrect. What is usually recomennded is using generalized cross-validation chosen by minimizing the empirical prediction error and thus varying the number of principal components to arrive at the optimal quantity using a data-driven procedure.

Functional Time Series Estimation from Data

Suppose that we have a functional dataset, denoted by $\mathbb{X}_{np}$ i.e. we have n curves over p gridpoints in an interval. We can form the functional principal components and test

whether or not the first few of them are cumulatively accounting for over 90% of the variance. If we reject the null hypothesis of *VERARH*, then we are dealing with ARH(1) process. Furthermore, we can test all the principal components and record the ones that reject the null hypothesis. Then we have come up with a decomposition consisting of scalar AR(1) processes and we could use it for forecasting because, according to Theorem 4, the rest would be just white noise which we do not want to incorporate into the predictions. In this section, we will compare the performance of our method and the generalized empirical prediction error crossvalidation procedure, suggested in Bosq (2000) chapter 9 and Besse et al. (2000). Next we will introduce some notation and the quantities we need to calculate in order to do the forecasting experiment.

Let us have our discretized functional data in a matrix form $\mathbb{X}_{n\times p}$ and they centred i.e.

$$\mathbb{E}[\mathbb{X}_{1:n,p}] = 0 \qquad \forall p \tag{7.8}$$

Then the covariance matrix of our $n \times p$ functional data with dimensions $p \times p$ is:

$$\boldsymbol{V_0} = \mathbb{E}[\mathbb{X}_{np}^T \mathbb{X}_{np}]. \tag{7.9}$$

The covariance matrix of lag 1 is:

$$\boldsymbol{V_1} = \mathbb{E}[\mathbb{X}_{n+1,p}^T \mathbb{X}_{np}]. \tag{7.10}$$

Based on the model (3.14) and Yule-Walker equations, the estimator of ρ from data is:

$$\hat{\boldsymbol{\rho}} = \boldsymbol{V_1 V_0^{-1}}. \tag{7.11}$$

The hat on ρ is because we are dealing with a discretized version of the model (3.14) over a grid of points and (7.11) is an estimator using all the gridpoints. Next, we will define our empirical estimators following Hörmann and Kokoszka (2012)

Now we want to apply the criterion from the procedure described in section (7.1.3) and have a functional estimator $\widetilde{\boldsymbol{\rho}}$. There are two possibilities for it. In both of them the matrix $\boldsymbol{V_0^{-1}}$ will be replaced by the approximate inverse using the number of chosen eigenvalues k in section (7.1.3) defined by:

Definition 17. *Let* $\widetilde{\boldsymbol{V_0^{-1}}}$ *approximated lag zero covariance given by*

$$\widetilde{\boldsymbol{V_0^{-1}}} = E\Lambda_k^{-1}E^T,$$

where E *is the matrix of the* k *AR(1)-like eigenvectors of* $\boldsymbol{V_0}$ *selected by section 7.1.3 and* Λ_k^{-1} *is the diagonal matrix with* k *nonzero elements, each equal to* $(1/\lambda_j)$ *for* $j = 1 \dots k$ *that are eigenvalues of* $\boldsymbol{V_0}$ *and* k *is, as previously, chosen in section (7.1.3)*

Definition 18. *Let* $\widetilde{\boldsymbol{V_1}}$ *be the approximated lag one covariance given by*

$$\widetilde{\boldsymbol{V_1}} = E'\Lambda_k'E'^T,$$

where E' *is the matrix of the eigenvectors of the symmetrized lag one covariance i.e.* $\boldsymbol{V_1^{sym}} = 1/2(\boldsymbol{V_1}+\boldsymbol{V_1^T})$ *and* Λ_k' *is the diagonal matrix with elements* (λ_j') *for* $j = 1, \dots, k$ *that are eigenvalues of* $\boldsymbol{V_1^{sym}}$ *and* k *is, as previously, chosen in 7.1.3.*

Now two possibilities for our functional estimator are suggested. The first one, which we will call *standard*, is:

Definition 19. *Definition of the Standard Estimator,* $\widetilde{\boldsymbol{\rho_1}}$*, is*

$$\widetilde{\boldsymbol{\rho_1}} = \boldsymbol{V_1}\widetilde{\boldsymbol{V_0^{-1}}}. \tag{7.12}$$

The second option, which we will call *smoothed*, uses the approximate version of the lag one covariance given in Definition 18

Definition 20. *Definition of the Smoothed Estimator,* $\widetilde{\boldsymbol{\rho_2}}$*, is*

$$\widetilde{\boldsymbol{\rho_2}} = \widetilde{\boldsymbol{V_1}}\widetilde{\boldsymbol{V_0^{-1}}}. \tag{7.13}$$

Then, the predictions are given by:

$$\widehat{\mathbb{X}_{n+1}} = \widetilde{\boldsymbol{\rho_1}}(\mathbb{X}_n) \qquad and \qquad \widehat{\mathbb{X}_{n+1}} = \widetilde{\boldsymbol{\rho_2}}(\mathbb{X}_n) \tag{7.14}$$

Prediction Equations

So, to summarize, in order to do prediction for ARH(1), using the proposed method in section 7.1.3 for selecting the number of principal components, we have to do three steps:

1. Select the number of principal components which have rejected the null hypothesis by using the algorithm in section (7.1.3)

2. Use the result from step 1 in calculating the estimators in (7.12) and (7.13)

3. Use the calculated estimators from step 2 in constructing the actual predictions, using (7.14)

Generalized Cross-validation (CV) by minimizing Empirical Prediction Error

As already emphasized, a crucial quantity in defining any functional estimator, is the number of principal components (k) to retain in the expansion of the covariance/correlation. The procedure, recommended by Bosq (2000); Ferraty and Vieu (2006); Besse et al. (2000) is by minimization of empirical prediction error function by generalized cross-validation (CV). The CV measure estimates the true prediction error and thus is most

natural when forecasting any kind of time series stochastic process. The criterion to minimize, then, is the following function of k:

$$\widehat{\Delta_n}(k) = \frac{1}{n - n_0} \sum_{v=n_0}^{n-1} \|\hat{\boldsymbol{\rho}}_v^{(k)}(\mathbb{X}_v) - \mathbb{X}_{v+1}\|^2 \tag{7.15}$$

where:

- $\hat{\boldsymbol{\rho}}_v^{(k)}$ is an estimator such as (7.11), (7.12) or (7.13) and the superscript k is indicating the number of principal components used to calculate it.

- n_0 is a quantity which defines the proportion between training and validation set that is usually chosen such that the first 80% of data are used for training and last 20% used for validation

- The norm in the sum could be L_1, L_2, L_∞ or any other applicable functional/matrix norm.

Then the optimal k for $\hat{\boldsymbol{\rho}}_v^{(k)}$ will be chosen as:

$$k = \underset{1 \leq k \leq p}{\arg\min} = \widehat{\Delta_n}(k) \tag{7.16}$$

Our method (*VERARH*) will be compared with the generalised CV approach, with respect to forecasting simulated ARH(1) data in section 7.4.

The simulation routine that we will use for generating data for the ARH(1) forecasting experiment in section 7.4 is from the `R` package `far` from Damon and Guillas (2010). Their routine `simulate.far.wiener` is based on Theorem 4 and it defines an ARH(1) with Wiener noise. Details of the routine as well as the setup of the simulation and the results of the application of *VERARH* for order verification of ARH(1) appear in appendix section **??**. The results are very good. Moreover, the idea is to numerically show that the theoretical foundations of our method are applicable in practice.

In this section we will use the performances of the two estimators (7.12) and (7.13) to construct predictions (7.14) in order to assess how our method *VERARH*, proposed in section(7.1.3), performs for forecasting in comparison to the generalised CV method from 7.15, proposed by Bosq (2000).

Obviously, the GCV method can be used to any time series. However, if our method performs at par with it, this is another layer of validation of our assertion that our method is good for *verification of the order* of an ARH(1) process.

7.4.2 Results from Forecasting Simulated Data

We generated $n = 2000$ curves over $p = 100$ gridpoints using the `simul.far.wiener` routine from the `far` package from Damon and Guillas (2010). We then used the following procedure:

1. The first n_1 number of curves will be used to define the training sample.

2. The lag 0 and 1 covariance/correlation will be calculated with data up to $n_1 - 1$ inclusive i.e. $\boldsymbol{V_0} = \mathbb{E}[\mathbb{X}^T_{(n_1-1)p}\mathbb{X}_{(n_1-1)p}]$
   $\boldsymbol{V_1} = \mathbb{E}[\mathbb{X}^T_{(n_1-1)p}\mathbb{X}_{(n_1-2)p}]$

3. The forecast will be calculated by $\widehat{\mathcal{X}_{n_1+1}} = \widetilde{\boldsymbol{\rho_1}}(\mathcal{X}_{n_1})$ and similarly for $\widetilde{\boldsymbol{\rho_2}}$. The settings for the CV method in *far* package are the default which is 80% training and 20 % validation.

4. The above step will be repeated 100 times by sliding the curves with one step ahead each iteration.

5. We will vary n_1 by 100 to illustrate performance accordingly for different portions of the data, each with a sliding forecast horizon of 100 curves.

In table 7.4 the last two columns were calculated in terms of MAPE(mean absolute percent error) and the comparisons were made curvewise i.e. for the 100 one-step ahead

n_1	k by CV	k by *VERARH*	Relative time ρ_1 is better	Relative time ρ_2 is better
1000	28	88	45 %	48%
1100	18	88	48 %	48%
1200	17	87	48 %	59%
1300	16	88	48 %	40%
1400	50	88	42 %	47%
1500	43	87	47 %	42%
1600	33	88	53 %	52%
1700	35	88	49 %	55%
1800	18	88	49 %	45%
1900	18	89	49 %	46%

Table 7.4: Table of the forecast comparisons using our method versus the CV method for defining the number of eigenvalues to retain in the expansion of the lag 0 covariance for ρ_1 and symmetrized lag 1 autocorrelation for ρ_2 respectively

forecasts the 100 errors from each method — CV and ours — were compared. We show percentages, but as they were 100 curves forecasted with each subset of data, they are also absolute results. The average number of times $\boldsymbol{\rho_1}$ was better is 47.80 % and for $\boldsymbol{\rho_2}$ this is 48.20 %. Our method's performance is slightly worse than the CV method. Another interesting thing to notice in table 7.1 is how the number of eigenvalues chosen, k, varies by each method. We may notice that the generalised CV method is very sensitive and changes are frequent. However, our method is quite robust with respect to k, in the sense that the amount is almost constant, which is a good thing, especially if we have to have an automatic forecasting system.

Furthermore, if we look at a typical picture which shows the two errors for the 100 steps, things look interesting. Firstly, let us have a look at the standard predictor (7.12) — On Fig. (7.7) we can see that our predictor (in green) is doing better at the extreme cases, when the MAPE is more than 1000%, Then, let us have a look at the smoothed predictor (7.13) on Fig. (7.8), it looks smoother than the standard one, and again, our method is slightly better at the extremes.

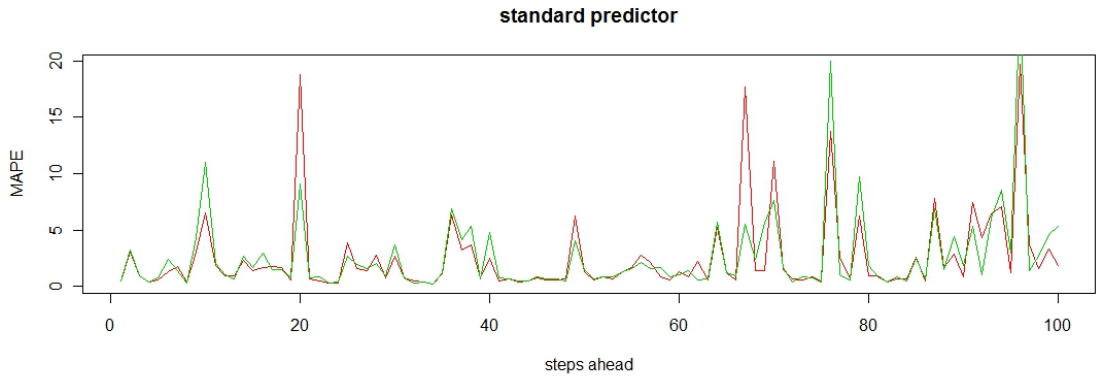

Figure 7.7: Plot of typical MAPE distribution over the 100 curves forecast horizon for the $\widetilde{\boldsymbol{\rho_1}}$ predictor. Green — our method for choosing k, Red — generalised CV

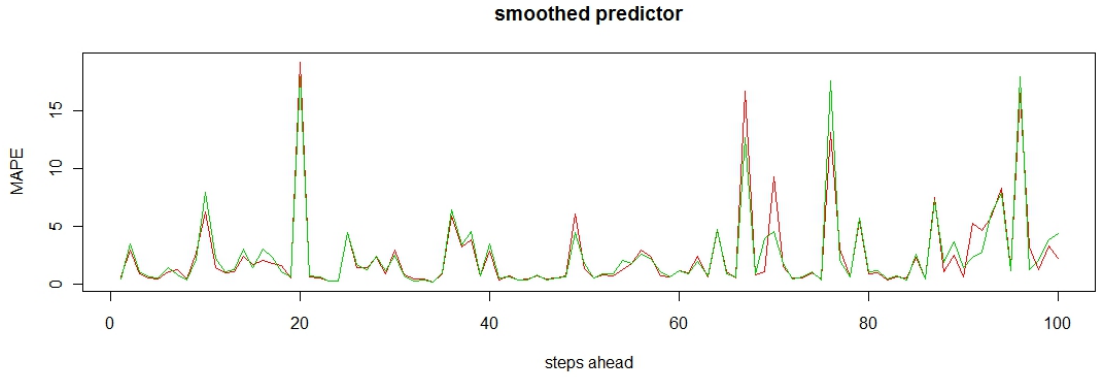

Figure 7.8: Plot of typical MAPE distribution over the 100 curves forecast horizon for the $\widetilde{\boldsymbol{\rho_2}}$ predictor. Green — our method for choosing k, Red — generalised CV

Summary and Implications of Forecasting Results

When used for prediction, our method for choosing the number of principal components k, performs at par with the standard empirical error GCV. In detail, it performs slightly worse for small errors and slightly better for extreme errors. This is a good indication that also the method is good for detection, bearing in mind the fact that the number of the principal components does not change almost at all when using different subsets of the data.

7.5 Conclusion and Further Work

A new technique was proposed for verification of ARH process for functional time series. The method performed well in simulation studies for ARH(1) and ARH(2). It could also be extended to ARH process of any reasonable order. The procedure was also applied to a real functional dataset with practical importance. An interesting question in this direction would be to do a seasonal ARH(p) in the fashion of seasonal ARIMA. On the other hand, the seasonal correlation structure might be varying in magnitude, so this gives rise to the question of time-varying ARH process.

Furthermore, our method gives a decomposition which could also be used for forecasting. Its performance was very close to the one of the GCV criterion, which is a further level of validation that our method is good for verifying the order of an ARH process.

Chapter 8

Conclusions and Innovations

White noise testing

In this thesis we used wavelets in order to construct tests for white noise for both univariate time series and two-dimensonal(spatial data). Our results from univariate white noise testing confirmed the usability of our tests and showed situations in which they are better than established tests and, naturally, vice versa. Furthermore, in chapters 4 and 5, we showed that our tests are applicable to a range of alternative hypotheses. This is achieved by the different types of wavelets that are used. We also developed a theoretical power function for the general wavelet test *genwwn* against an alternative of an ARMA process. The power function is a very useful tool in practice since it gives answer to the question "How large does my sample have to be in order to detect an alternative hypothesis of class ARMA(p, q)?". Moreover, many tests in the literature do not have a theoretical power function. We also showed a range of real data examples. Another important feature of our univariate wavelet white noise tests is that they do not require any manual tuning parameters to be determined by the user. This is in contrast to many of the available tests in the literature.

A wavelet test based on a single coarsest scale wavelet coefficient was shown to be equivalent to a weighted sum of the odd-lagged autocorrelation coefficients which explores a contrast beween high and low frequencies.The $d00$ test is an interesting area for

further research itself, since a similar test can be created from even-lagged autocorrelations. The test performed very well against alternatives such as 'hidden periodicity' and small magnitude autocorrelation, which are subject of contemporary reseach. We also showed a range of real data examples and our tests performed well.

Although approximate, our result for the distribution of the Haar wavelet coefficients of the periodogram (from Proposition 2, equation (4.6)) seems to have very good perfomarmance in the simulation study. An area of further methodological research for the *hwwntest* would be to analyse it with respect to an arbitrary standard deviation σ (we used $\sigma^2 = 1$ for the proof) since this would have an implication on the distribution. For instance we might have white noise with changing regimes i.e. $\sigma(t)$, which would be an interesting topic for further work. Last, but not least, we have created an `R` package `hwwntest` (Savchev and Nason, 2015) implementing the tests and providing functions for simulations and theoretical power calculations.

We extended the univariate *hwwntest* white noise test to two-dimensional spatial data in chapter 5. To the best of our knowledge, we are not aware of other spatial autocorrelation tests based on wavelets. We compared our *2d hwwntest* with the classical spatial autocorrelation test of Moran in a simulation study. Our test performed reasonably, though it has some deficiencies — low power for a small size dataset. This has to do with the fact that due to the setup of the two-dimensional orthogonal Haar wavelet transform, we had to replicate the periodogram which induced more tests for each wavelet coefficient, thus requiring more multiple comparisons to adjust for. An interesting avenue for further work would be to revise our test by selecting the wavelet coefficients that correspond to the 'first half' of the two-dimensional periodogram or correct the nominal size of the test. Another direction would be to use a non-decimated wavelet transform or even wavelet packets, since spatial autocorrelation is a phenomenon that can be expressed in many different ways. Wavelet methods have future in spatial statistics and our test is but one initial step in this endeavour. Last but not least, we also did analysis of one of the well-known

spatial datasets, the Mercer and Hall(1911) Wheat data, and our test detected correctly that there is a spatial trend in the data.

ARH order verification

In chapter 6, based on an established theoretical result for ARH(1) process existence, we developed a testing procedure that can be applied for checking if a functional time series dataset follows an ARH(1) process. We validated our methodology by a simulation study. Furthermore, since there are more supporting theoretical results, we extended our methodology by a multistaging procedure for verification of the order of an ARH(p) process. We did a comparison with one of the few available such procedures in the literature, for an ARH(2), and our results were good. Moreover, it turned out that our methodology for order verification of ARH(1) could also be used for forecasting. We performed an experiment with simulated data against a well-known generalised CV procedure and our results were on par with it. Given the fact that our procedure is simpler than generalised CV, we believe it is worth further research. For instance, many functional time series datasets that come from observation of continuous processes, such as electricity load or magnetic field measurement, could have strong seasonal characteristics similar to seasonal ARIMA processes. Thus, it would be interesting to develop a methodology for seasonal ARH(p) estimation. Moreover, seasonality gives rise to time varying parameters, so a time-varying ARH(1) process would be an important direction to pursue.

***To sum up, the main conclusions of our work are*:**

1. White noise testing is a tricky problem since the possible space of alternative hypotheses is vast and there are no universal tests which can cover all of them with uniform power.

2. There is a much room for development of white noise tests and using wavelets in the spectral domain provides a flexible way to design different tests .

3. There exists theoretical basis from which a methodology can be derived for testing

of ARH processes of any order.

4. The same methodology could be used for forecasting of ARH processes.

The main innovations of this work are:

1. Developing wavelet-based tests for white noise in the spectral domain as opposed to contemporary tests which are mostly heavily-modified Box-Pierce-Ljung type tests, each with a few tuning parameters.

2. Developing a theoretical power function for the *genwwn* test.

3. The different wavelet tests serve different purposes and thus with one paradigm we cover different subsets of alternative hypotheses e.g. 'hidden periodicity' or small magnitude autocorrelation. Moreover, our tests do not require any tuning parameters to be provided or guessed by the user.

4. Developing a wavelet test for spatial autocorrelation, without specifying a spatial lag, i.e. number of neighbours.

5. Implementation of the wavelet tests in a separate `R` package `hwwntest`.

6. Suggesting a feasible methodology for order verification of ARH processes and facilitating its interface with the mathematical foundations of ARH processes.

7. Using this methodology also for forecasting and showing a fair performance against a technique (generalised CV) which is explicitly suited for forecasting of any time series process.

Bibliography

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach, *J. R. Statist. Soc.* B, **60**, 725–749.

Ahdesmaki, M., Fokianos, K., and Strimmer., K. (2012) *GeneCycle: Identification of Periodically Expressed Genes*, r package version 1.1.2.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J. (2010) Modelling and forecasting daily electricity load curves: a hybrid approach, *Proceedings of the Nineteenth International Conference on Computational Statistics (COMPSTAT)*.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2013) Clustering functional data using wavelets, *Int. J. Wavelets Multiresolut Inf. Process*, **11**.

Barber, S., Nason, G. P., and Silverman, B. W. (2002) Posterior probability intervals for wavelet thresholding, *J. R. Statist. Soc.* B, **64**, 189–205.

Bartlett, M. (1954) Probleèmes de l'analyse spectrale des séries temporelles stationnaires, *Publ. Inst. Statist. Univ. Paris*, **3**, 119–34.

Bartlett, M. (1955) *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*, Cambridge University Press, Cambridge.

Bartlett, M. S. (1950) Periodogram analysis and continuous spectra, *Biometrika*, **37**, 1–16.

Baxter, C. and Rennie, A. (1996) *Financial Calculus: An Introduction to Derivative Pricing*, Cambridge University Press.

Benjamini, Y. and Hochberg, Y. (1995a) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc.* B, **57**, 289–300.

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

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, Series B*, **36**, 191–225.

Besag, J. (1977) Errors-in-variables estimation for gaussian lattice schemes, *Journal of the Royal Statistical Society, Series B*, **39**, 73–78.

Besse, P. and Ramsay, J. (1986) Principal components analysis of sampled functions, *PSYCHOMETRIKA*, **51**, 285–311.

Besse, P., Cardot, H., and Stephenson, D. (2000) Autoregressive forecasting of some functional climatic variations, *Scandinavian Journal of Statistics*, **27**, 673–687.

Bivand, R. and Piras, G. (2015) Comparing implementations of estimation methods for spatial econometrics, *Journal of Statistical Software*, **63**, 1–36.

Bivand, R., Hauke, J., and Kossowski, T. (2013) Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods, *Geographical Analysis*, **45**, 150–179.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307 – 327.

Bosq, D. (1991) Modelization, nonparametric estimation and prediction for continuous time processes., in G. Roussas, ed., *Nonparametric functional estimation and related topics*, NATO ASI Series, pp. 509–529, NATO.

Bosq, D. (2000) *Linear Process in Function Spaces*, Springer Series in Statistics, Springer.

Box, G. E. P. and Pierce, D. A. (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *Journal of the American Statistical Association*, **65**, 1509–1526.

Brilliger, D. (2001) *Time Series: Data Analysis and Theory*, Philadelphia: Society for Industrial Mathematics.

Brillinger, D. (1969) Asymptotic properties of spectral estimates of second order, *Biometrika*, **56**, 375–390.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, Springer, New York.

Burridge, P. (1980) On the cliff-ord test for spatial correlation, *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**, 107–108.

Chatfield, C. (1996) *The Analysis of Time Series: An Introduction*, Chapman and Hall/CRC, London, fifth edition.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013) Modelling and forecasting daily electricity load curves: a hybrid approach, *Journal of the American Statistical Association*, **108**, 7–21.

Cliff, A. and Ord, K. (1972) Testing for spatial autocorrelation among regression residuals, *Geographical Analysis*, **4**, 267–284.

Coifman, R. R. and Donoho, D. L. (1995) Translation-invariant de-noising, in A. Antoniadis and G. Oppenheim, eds., *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pp. 125–150, Springer-Verlag, New York.

Cooley, J. and Tukey, J. W. (1965) An algorithm for the machine calculation of complex fourier series, *Mathematics of Computation*, **19**, 297–301.

Cressie, N. (1985) A geostatistical analysis for mercer and hall wheat data, *Bulleting of the International Statistical Institute*, **51**, 277–78.

Cressie, N. (1993) *Statistics for Spatial Data*, Wiley: Chichester.

Damon, J. and Guillas, S. (2005) Estimation and simulation of autoregressivehilbertian processes with exogenous variables, *Statistical Inference for Stochastic Processes*, **8**, 185–204.

Damon, J. and Guillas, S. (2010) Far package help, *CRAN*.

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

Deo, R. (2000) Spectral tests of the martingale hypothesis under conditional heteroscedasticity, *Journal of Econometrics*, **99**, 291–315.

Donoho, D. L. (1993a) Nonlinear wavelet methods of recovery for signals, densities, and spectra from indirect and noisy data, in *Proceedings of Symposia in Applied Mathematics*, volume 47, American Mathematical Society, Providence: RI.

Donoho, D. L. (1993b) Unconditional bases are optimal bases for data compression and statistical estimation, *App. Comp. Harm. Anal.*, **1**, 100–115.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal denoising in an orthonormal basis chosen from a library of bases, *Compt. Rend. Acad. Sci. Paris A*, **319**, 1317–1322.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995) Wavelet shrinkage: Asymptopia? (with discussion), *J. R. Statist. Soc.* B, **57**, 301–369.

Doob, J. (1953) *Stochastic Processes*, Wiley, New York.

Durbin, J. (1969) Testing for serial correlation in regression analysis based on the periodogram of least-squares residuals, *Biometrika*, **56**, 1–56.

Durbin, J. (1970) Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables, *Econometrica*, **38**, 410–421.

Durbin, J. and Watson, G. S. (1950) Testing for serial correlation in least squares regression.i, *Biometrika*, **37**, 409–427.

Durbin, J. and Watson, G. S. (1951) Testing for serial correlation in least squares regression.ii, *Biometrika*, **38**, 159–177.

Durbin, J. and Watson, G. S. (1971) Testing for serial correlation in least squares regression.iii, *Biometrika*, **58**, 1–19.

Engle, R. (1984) Wald, likelihood ratio, and lagrange multiplier tests in econometrics, in Z. Griliches† and M. D. Intriligator, eds., *Handbook of Econometrics*, volume 2, chapter 13, pp. 775–826, Elsevier, 1 edition.

Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, *Econometrica*, **50**, pp. 987–1007.

Escanciano, J. and Lobato, I. (2009) An automatic portmanteau test for serial correlation, *Journal of Econometrics*, **151**, 140–149.

Ferraty, F. and Vieu, P. (2006) *NonParametric Functional Data Analysis. Theory and Practice*, Springer Series in Statistics.

Fisher, R. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika*, **10**, 507–529.

Fisher, R. (1929) Tests of significance in harmonic analysis, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **125**, 54–59.

Geary, R. C. (1954) The contiguity ratio and statistical mapping, *The Incorporated Statistician*, **5**, 115–127+129–146.

Gneiting, T. (1997) Normal scale mixtures and dual probability densities., *J. Statist. Comp. Simul.*, **59**, 375–384.

Grenander, U. (1981) *Abstract Inference*, Wiley.

Grenander, U. and Rosenblatt, M. (1957) *Statistical Analysis of Stationary Time Series*, Wiley, New York.

Guay, A., Guerre, E., and Štěpána, L. (2013) Robust adaptive rate-optimal testing for the white noise hypothesis, *Journal of Econometrics*, **176**, 134–145.

Hallin, M., Ingenbleek, J.-F., and Puri, M. L. (1987) Linear and quadratic serial rank tests for randomness against serial dependence, *Journal of Time Series Analysis*, **8**, 409–424.

Hatekar, N. (2010) *Principles of Econometrics: An Introduction(Usng R)*, SAGE Publications, New Delhi, India.

Herrick, D. R. M., Nason, G. P., and Silverman, B. W. (2001) Some new methods for wavelet density estimation, *Sankhyā* A, **63**, 391–411.

Hong, Y. (1996) Consistent testing for serial correlation of unknown form, *Econometrica*, **64**, 837–864.

Hong, Y. and Lee, Y. (2005) Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form, *Rev. Econ. Stud.*, **72**, 499–541.

Hörmann, S. and Kokoszka, P. (2012) Functional time series, *Handbook of Statistics: Time Series Analysis: Methods and Applications*, **30**.

Hyndman, R. (2014) Thoughts on the ljung-box test, *Blog post*.

Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise, *J. R. Statist. Soc.* B, **59**, 319–351.

Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G. (1999) Surface air temperature and its changes over the past 150 years, *Reviews of Geophysics*, **37**, 173–199.

Koen, C. (1990) Significance testing of periodogram ordinates, *Astrophys. J.*, **348**, 700–702.

Kokoszka, P. and Reimherr, M. (2013) Determining the order of the functional autoregressive model, *Journal of Time Series Analysis*, **34**, 116–129.

Kolmogorov, A. (1933) Sulla determinazione empirica di una legge di distribuzione, *G. Ist. Ital. Attuari*, pp. 83–91.

Kotz, S. and Johnson, N. (1992) *Breakthroguhs in statistics Volume II. Methodology and Distribution*, Springer Series in Statistics, Springer.

Kotz, S., Kozubowski, T., and Podgorski, K. (2001) *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser Basel.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014) The parable of google flu: Traps in big data analysis, *Science*, **343**, 1203–1205.

Li, H., Calder, C. A., and Cressie, N. (2007) Beyond moran's i: Testing for spatial dependence based on the spatial autoregressive model, *Geographical Analysis*, **39**, 357–375.

Ljung, G. and Box, G. (1978) On a measure of lack of fit in time series models, *Biometrika*, **65**, 297–303.

Lobato, I. and Velasco, C. (2004) A simple and general test for white noise, in *Econometric Society 2004 Latin American Meetings*, Econometric Society.

Lobato, I., Nankervis, J. C., and Savin, N. E. (2001) Testing for autocorrelation using a modified box-pierce q test, *International Economic Review*, **42**, pp. 187–205.

Lobato, I. N. (2001) Testing that a dependent process is uncorrelated, *Journal of the American Statistical Association*, **96**, 1066–1076.

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

Mallat, S. G. (1998) *A Wavelet Tour of Signal Processing*, Academic Press, San Diego.

McCoy, E., Walden, A., and Percival, D. (1998) Multitaper spectral estimation of power law processes, *IEEE Transactions on Signal Processing*, **46**, 655–668.

Mercer, W. B. and Hall, A. D. (1911) The experimental error of field trials, *Journal of Agricultural Science*, **4**, 107–132.

Meyer, Y. (1993) *Wavelets and Operators*, Cambridge University Press, Cambridge.

Moran, P. A. P. (1950) Notes on continuous stochastic phenomena, *Biometrika*, **37**, 17–23.

Nason, G. and Savchev, D. (2014a) Supplementary material for "white noise testing using wavelets", Technical Report 14:01, Statistics Group, University of Bristol.

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

Nason, G. P. (2013) A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series., *J. R. Statist. Soc.* B, **75**, 879–904.

Nason, G. P. and Savchev, D. (2014b) White noise testing using wavelets, *Stat*, **3**, 351–362.

Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S, *J. Comp. Graph. Stat.*, **3**, 163–191.

Nason, G. P. and von Sachs, R. (1999) Wavelets in time series analysis, *Phil. Trans. R. Soc. Lond. A*, **357**, 2511–2526.

Nason, G. P., von Sachs, R., and Kroisandt, G. (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum, *J. R. Statist. Soc.* B, **62**, 271–292.

Neumann, M. (1994) Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series, Statistics Research Report SRR 028-94, Australian National University, Canberra, Australia.

Neumann, M. (1996) Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series, *J. Time Ser. Anal.*, **17**, 601–633.

Newton, J. (1996) sts12: A periodogram-based test for white noise, *Stata Technical Bulletin*, **34**, 36–39, College Station TX: Stata Press.

Ord, K. (1975) Estimation methods for models of spatial interaction, *Journal of the American Statistical Association*, **70**, 120–126.

Pawitan, Y. (1996) Automatic estimation of the cross-spectrum of a bivariate time series, *Biometrika*, **83**, 419–432.

Percival, D. B. (1995) On estimation of the wavelet variance, *Biometrika*, **82**, 619–631.

Percival, D. B. and Walden, A. T. (2000) *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge.

Priestley, M. B. (1983) *Spectral Analysis and Time Series*, Academic Press, London.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ramsay, J. (1996) Principal differential analysis: Data reduction by differential operators, *Journal of the Royal Statistical Society, Series B*, **58**, 233–243.

Ramsay, J. and Dalzell, C. (1991) Some tools for functional data analysis, *Journal of the Royal Statistical Society, Series B*, **53**, 539–572.

Ramsay, J. and Silverman, B. (2002) *Functional Data Analysis*, Springer Series in Statistics.

Rao, T. S., Das, S., and Boshnakov, G. N. (2014) A frequency domain approach for the estimation of parameters of spatio-temporal stationary random processes, *Journal of Time Series Analysis*, **35**, 357–377.

Reschenhofer, E. (1989) Adaptive test for white noise, *Biometrika*, **76**, 629–632.

Rice, J. and Silverman, B. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society, Series B*, **53**, 233–243.

Ripley, B. (1981) *Spatial Statistics*, Wiley, New York.

Robinson, P. M. and Rossi, F. (2014) Improved lagrange multiplier tests in spatial autoregressions, *The Econometrics Journal*, **17**, 139–164.

Rudzkis, R., Saulis, L., and Statulevičius, V. (1978) A general lemma on probabilities of large deviations, *Lithuanian Math. J.*, **18**, 226–238.

Samuelson, P. (1965) Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review*, **6**, 41–49.

Savchev, D. and Nason, G. (2015) *hwwntest: Tests of White Noise using Wavelets*, R package version 1.3.

Schuster, A. (1898) On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena, *Terrestrial Magnetism*, **3**, 13–41.

Silvey, S. D. (1959) The lagrangian multiplier test, *The Annals of Mathematical Statistics*, **30**, 389–407.

Smirnov, N. (1948) Table for estimating the goodness of fit of empirical distributions, *Ann. Math. Statist.*, **19**, 279–281.

Stuart, A. and Ord, J. K. (1994) *Kendall's Advanced Theory of Statistics: Distribution Theory*, volume 1, Arnold, London.

Tukey, J. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*, Wiley, New York.

von Sachs, R. and Neumann, M. H. (2000) A wavelet-based test for stationarity, *J. Time Ser. Anal.*, **21**, 597–613.

Wahba, G. (1980) Automatic smoothing of the log periodogram, *J. Am. Statist. Ass.*, **75**, 122–132.

Walden, A., Percival, D., and McCoy, E. (1998) Spectrum estimation by wavelet thresholding of multitaper estimators, *IEEE Transactions on Signal Processing*, **46**, 3153–3165.

Walter, G. G. (1994) *Wavelets and Other Orthogonal Systems with Applications*, Chapman and Hall, Boca Raton.

Wasserman, L. (2005) *All of Nonparametric Statistics*, Springer, New York.

Whitcher, B. (2015) *waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing*, r package version 1.7.5.

Whittle, P. (1954) On stationary processes in the plane, *Biometrika*, **41**, 434–449.

Wuertz, D., with contribution from Michal Miklovic, Y. C., Boudt, C., Chausse, P., and others (2016) *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*, r package version 3010.82.1.

Appendix A

Proofs, ARH simulations and software

A.1 Proof of Proposition 2

Recall that the characteristic function of an exponential random variable with parameter λ is $\phi(t) = (1 - it/\lambda)^{-1}$ where $\sqrt{-1} = i$. Hence, if X, Y are independent exponential distributed variables then the characteristic function of $W = 2^{-1/2}(X - Y)$, the finest scale wavelet coefficient, is given by

$$\begin{aligned} \phi_W(t) &= \phi_W(u) = \phi_X(u)\phi_Y(-u) = (1 - iu/\lambda)^{-1}(1 + iu/\lambda)^{-1} \\ &= \lambda^2/(\lambda^2 + u^2) = \lambda^2/(\lambda^2 + t^2/2), \end{aligned} \tag{A.1}$$

which is the characteristic function of the double exponential or Laplace distribution, where $t = \sqrt{2}u$.

More generally, for scale $\ell = J - j$ of the Haar wavelet transform the wavelet coefficient $\hat{v}_{j,k}$ is the difference, W, of two random variables, X, Y each of which are $2^{\ell/2}$ times the sum of $2^{\ell-1}$ exponential random variables. Hence, the characteristic function of

$\hat{v}_{j,k}$ is given by

$$\begin{aligned}\phi_W(t) &= \phi_X(u)\phi_Y(-u) = (1-iu/\lambda)^{-2^{\ell-1}}(1+iu/\lambda)^{-2^{\ell-1}} \\ &= \left\{\lambda^2/(\lambda^2+u^2)\right\}^{2^{\ell-1}} \\ &= \left\{\lambda^2/(\lambda^2+t^2/2^{\ell})\right\}^{2^{\ell-1}}, \end{aligned} \tag{A.2}$$

where $t = 2^{\ell/2}u$. For example, for the finest scale wavelet coefficient $j = J-1$ which implies $\ell = 1$ and formula (A.2) coincides with (A.1).

Formula (A.2) is precisely the probability density function of a Student's t-distribution whose characteristic function can be found in Stuart and Ord (1994, Ex. 3.13). Hence, due to the duality property of Fourier transforms the probability density of $\hat{v}_{j,k}$ is given by (4.6) and is of a similar form to the characteristic function of a Student's t-distribution. We refer to the density in (4.6) as Macdonald's distribution as it is essentially Macdonald's special function, a cylinder function or a modified Bessel function of the third kind, see also Gneiting (1997). The distribution for $\hat{v}_{j,k}$ in (4.6) has mean zero and variance one.

A.2 Proof of Proposition 3

Our wavelet coefficients $\hat{v}_{j,k}$ are a discrete version of the integral version of the wavelet coefficient defined by (Neumann, 1996, 2.3)

$$\tilde{v}_{j,k} = \int \psi_{j,k}(\omega) I_T(\omega)\, d\omega, \tag{A.3}$$

where $I_T(\omega)$ is the periodogram in (4.1). Neumann (1994) establishes that the asymptotic behaviour of both is the same.

Following Neumann (1996) let $\sigma^2_{j,k}$ denote $\mathrm{var}(\tilde{v}_{j,k})$, define

$$\mathcal{J} = \mathcal{J}(T) = \{(j,k) |\, 2^j \leq CT^{1-\delta}, k \in I_j\},$$

where $C < \infty$, $0 < \delta \leq 1/3$ are arbitrary constants and $I_j = \{1, \ldots, 2^j\}$.

This range is not restrictive to any practical dataset size T. Let us show this by an example:

$$2^j \;\leq\; CT^{1-\delta} \iff j \leq (1-\delta)J \iff \delta J \leq 1 \tag{A.4}$$

where we have taken $C = 1$, $J = log_2 T$ and $j = J - 1$ for the finest scale. Since δ can be arbitrarily small, let $\delta = 0.05$ and $T = 1024$, thus $J = 10$ and $j = 9$ then $\delta J = 0.5 \leq 1$. For scales coarser than $J - 1$ the right hand side of the last inequality in (A.4) can only increase.

Define $\Delta_\gamma = o\{\Delta^{1/(3+4\gamma)}\}$ where $\Delta = T^{\delta/2}(\log T)^{-1}$ and γ is specified in Assumption 2. Further define

$$\sigma_T = \max\{\max_{(j,k)\in\mathcal{J}}(\sigma_{j,k}, C_0 T^{-1/2})\}$$

for some fixed $C_0 > 0$ and $\theta_{j,k} \sim N(0, \sigma_T^2 - \sigma_{j,k}^2)$. Then the following theorem, which relies on the large deviation result Lemma 1 from Rudzkis et al. (1978), establishes the asymptotic normality of the $\tilde{v}_{j,k}$.

Theorem 4.2 Neumann (1996). Assume Assumptions 2–4. Then

$$\frac{\mathbb{P}[\pm\{(\tilde{v}_{j,k} + \theta_{j,k}) - v_{j,k}\}/\sigma_T \geq x]}{1 - \Phi(x)} \to 1$$

holds uniformly in $(j,k) \in \mathcal{J}$, $-\infty < x \leq \Delta_\gamma$.

Neumann (1996) Proposition 3.1(ii) shows that

$$\sigma_{j,k}^2 = 2\pi T^{-1} \int_\Pi \psi_{j,k}(\omega)\{\psi_{j,k}(\omega) + \psi_{j,k}(-\omega)\}|f(\omega)|^2\, d\omega + o(T^{-1}) + \mathcal{O}(T^{-1}2^{-j}),$$

for our situation and our wavelets this reduces to $\sigma_{j,k}^2 = 2\pi\sigma_X^{-2}T^{-1}\int_\Pi \psi_{j,k}^2(\omega) f^2(\omega)\, d\omega$.

Under H_0 the spectrum is constant $f(\omega) = (2\pi)^{-1}\sigma_X^2$ for $\omega \in \Pi$. Hence, under the

asymptotic regime above, the mean of $\tilde{v}_{j,k}$ is zero from (4.3) since the wavelet coefficients of a constant function are zero. The variance of $\tilde{v}_{j,k}$ is given by $\sigma^2_{j,k} = T^{-1}$ since $||\psi_{j,k}(\omega)||^2 = 1$. For the discrete coefficients the normalization is different by $T^{-1/2}$ so, asymptotically $\hat{v}_{j,k} \sim N(0,1)$. (Compare, for example, with the Haar wavelet test where the null Macdonald distribution in (4.6) has mean zero and variance one and asymptotically tends to the normal distribution).

Under H_A $\tilde{v}_{j,k} \sim N(v_{j,k}, \sigma^2_{j,k})$ asymptotically.

A.3 Proof of Approximation 4

Under H_0 the $\hat{v}_{j,k} \sim N(0,1)$ from Proposition 3 and assume that there are N_c coefficients to test. The nominal size of the test is α and the corrected Bonferroni size is $\alpha_c = N_c^{-1}\alpha$. Let the p-value of the (j,k)th test be $p_{(j,k)}$. We will reject H_0 if $\min_{(j,k)\in N_c} p_{j,k} < \alpha_c$. Now, for a given (j,k),

$$p_{(j,k)} < \alpha_c \equiv 2\{1 - \Phi(|\hat{v}_{j,k}|)\} < \alpha_c \equiv |\hat{v}_{j,k}| > \Phi^{-1}(1 - \alpha_c/2).$$

So, define the critical value for the test to be $C_{\alpha_c} = \Phi^{-1}(1 - \alpha_c/2)$.

For the power we now assume H_A is true and hence $\hat{v}_{j,k} \sim N(v_{j,k}, \sigma^2_{j,k})$ asymptotically. Hence the power function is:

$$\begin{aligned}
\mathbb{P}\{\text{Reject } H_0 | f(\omega)\} &= \mathbb{P}\left(C_{\alpha_c} < \max_{(j,k)\in\mathcal{I}_T} |\hat{v}_{j,k}|\right) \\
&= 1 - \mathbb{P}\left(\max_{(j,k)\in\mathcal{I}_T} \hat{v}_{j,k} \leq C_{\alpha_c}\right) \\
&= 1 - \prod_{(j,k)\in\mathcal{I}_T} \mathbb{P}(|\hat{v}_{j,k}| \leq C_{\alpha_c}) \\
&= 1 - \prod_{(j,k)\in\mathcal{I}_T} \left\{\Phi_{\eta_{j,k}}(C_{\alpha_c} - v_{j,k}) - \Phi_{\eta_{j,k}}(-C_{\alpha_c} - v_{j,k})\right\}, \quad \text{(A.5)}
\end{aligned}$$

A.4 Derivation of the $d_{0,0}$ index

Suppose we want to base a hypothesis test just on a single wavelet coefficient of the spectrum. We choose $d_{0,0}$. The point of this section is to show that there is a combination of autocorrelations which will furnish an identical test. We know that from the above theory that under H_0 this will be distributed $N(0,1)$ asymptotically, and since it is the coarsest scale coefficient the asymptotics should kick in quickly.

The coefficient is defined by

$$\begin{aligned} d_{0,0} &= \int_0^{\pi} f(\omega)\psi_{0,0}(\pi^{-1}\omega)\,d\omega && \text{(A.6)} \\ &= 2^{-1/2}\int_0^{\pi/2} f(\omega)\,d\omega - 2^{-1/2}\int_{\pi/2}^{\pi} f(\omega)\,d\omega, && \text{(A.7)} \end{aligned}$$

where $\psi_{j,k}(\omega)$ is the standard Haar wavelet on $[0,1]$. The spectrum can be expressed in terms of autocovariances by $f(\omega) = \pi^{-1}\sum_{k=-\infty}^{\infty}\gamma(k)\exp(-i\omega k)$ so

$$\begin{aligned} d_{0,0} &= (\sqrt{2}\pi)^{-1}\sum_{k=-\infty}^{\infty}\gamma(k)r_k && \text{(A.8)} \\ &= (\sqrt{2}\pi)^{-1}\left\{ r_0\gamma(0) + 2\sum_{k=1}^{\infty}(r_k + r_{-k})\gamma(k) \right\}, && \text{(A.9)} \end{aligned}$$

where

$$r_k = \int_0^{\pi/2} e^{-ik\omega}\,d\omega - \int_{\pi/2}^{\pi} e^{-ik\omega}\,d\omega \qquad \text{(A.10)}$$

It can be shown that (after some algebra) that $r_0 = 0$ and

$$r_k + r_{-k} = \frac{8\sin\pi k/4^2}{k}\sin\pi k/2. \qquad \text{(A.11)}$$

Then substituting (A.11) into (A.9) and only using the odd-indexed values (as the even

ones are zero) we get:

$$d_{0,0} = \frac{4}{\sqrt{2}\pi} \sum_{m=0}^{\infty} \gamma(2m+1)/(2m+1), \tag{A.12}$$

which is the formula coded into `d00.test`, except that we used autocorrelations there which normalizes (i.e. so you don't have to worry about the variance).

Similar formulae could be derived for $d_{1,0}, d_{1,1}$ and $d_{2,k}$ for $k = 0, 1, 2, 3$ and into further scales. Then by combining the results of these test you'd end up with a test like `hwwn.test` but only for these coarse scales. The point being that the autocovariances can be computed for *arbitrary* T easily. So, even though the test is wavelet based you don't use the wavelet transform to compute the values.

A.5 Proof of Proposition 5

Firstly, we need to derive the form of the distribution of the wavelet coefficients for the finest scale and then generalize it to any coarser scale l. However, each type of detail-level wavelet coefficient (diagonal, vertical and horizontal) has two additions and two subtractions in its definition, thus their distribution will be the same. Let us show it by a 4×4 illustration with the diagonal coefficients at the finest scale.

$$Data_{nn} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix},$$

Now, let us denote the first diagonal coefficient at the *finest scale* by D_f:

$$D_f = x_{11} - x_{12} - x_{21} + x_{22}, \tag{A.13}$$

Recall that the characteristic function of an exponential random variable with param-

eter λ is $\phi(t) = (1 - it/\lambda)^{-1}$ where $\sqrt{-1} = i$. Hence, if we want to check the distribution of the diagonal coefficient D_f from eq. (A.13), we set a random variable:

$$W_d = 2^{-1}(X_1 - X_2 - X_3 + X_4), \tag{A.14}$$

where X_1 to X_4 correspond to the independent and identically distributed Exponential random variables, corresponding to the first 4-block of data in the matrix from equation (A.13). Next we write down the characteristic function of this sum (A.14) of random variables :

$$\begin{aligned}
\phi_{W_d}(t) = \phi_W(u) &= \phi_{X_1}(u)\phi_{X_2}(-u)\phi_{X_3}(-u)\phi_{X_4}(u) \\
&= (1 - iu/\lambda)^{-1}(1 + iu/\lambda)^{-1}(1 + iu/\lambda)^{-1}(1 + -iu/\lambda)^{-1} \\
&= \{\lambda^2/(\lambda^2 + u^2)\}\{\lambda^2/(\lambda^2 + u^2)\} \\
&= \frac{\lambda^4}{(\lambda^2 + u^2)^2} \\
&= \frac{\lambda^4}{(\lambda^2 + t^2/4)^2}
\end{aligned} \tag{A.15}$$

which is the characteristic function of the sum/difference of two Laplace distributions or difference of Erlangs (Gamma) i.e. the Macdonald/Bessel, where $t = 2u$.

In order to derive the distribution for the coarser scales of the diagonal wavelet coefficients, we need firstly to think about the scaling functions coefficients in our four by four setup. The scaling function coefficients at the finest scale are defined as:

$$W_s = 2^{-1}(X_1 + X_2 + X_3 + X_4), \tag{A.16}$$

We note that the scaling coefficients at the finest scale are just sums of Exponential random variables i.e. Erlang/Gamma Thus their characteristic function is:

$$\phi_{W_s}(t) = \phi_{W_s}(u) \quad = \quad \phi^4_{X_1}(u) = \frac{\lambda^4}{(\lambda - iu)^4} \tag{A.17}$$

Now, the next coarser level wavelet diagonal coefficients are going to be the same linear combination as eq. (A.14), just we need to replace each of the X_i with a W_s from eq. (A.16):

$$\begin{aligned}
\phi_{Wd_{[finest+1]}} &= \phi^4_{X_1}(u)\phi^4_{X_1}(-u)\phi^4_{X_1}(-u)\phi^4_{X_1}(u) \\
&= \frac{\lambda^{16}}{(\lambda - iu)^4(\lambda + iu)^4(\lambda + iu)^4(\lambda - iu)^4} \\
&= \frac{\lambda^{16}}{(\lambda^2 + u^2)^4(\lambda^2 + u^2)^4} \\
&= \frac{\lambda^{16}}{(\lambda^2 + u^2)^8} = \frac{\lambda^{16}}{(\lambda^2 + t^2/4)^8}
\end{aligned} \tag{A.18}$$

which has the same form as (A.15) and is the characteristic function of the sum of eight Laplace random variables i.e. Macdonald/Bessel where $t = 2u$.

More generally, for scale $\ell = J - j$ of the 2D Haar wavelet transform, the wavelet coefficient W_{d_ℓ} is the difference, W, of two random variables, X, Y each of which are 2^ℓ times the sum of $2^{2\ell-1}$ exponential random variables. Hence, the characteristic function of W_{d_ℓ} is given by

$$\begin{aligned}
\phi_{W_{d_\ell}} &= \phi_X(u)\phi_Y(-u) = (1 - iu/\lambda)^{-2^{2\ell-1}}(1 + iu/\lambda)^{-2^{2\ell-1}} \\
&= \left\{\lambda^2/(\lambda^2 + u^2)\right\}^{2^{2\ell-1}} \\
&= \left\{\lambda^2/(\lambda^2 + t^2/2^{2\ell})\right\}^{2^{2\ell-1}},
\end{aligned} \tag{A.19}$$

where $t = 2^\ell u$. For example, for the finest scale diagonal wavelet coefficients $j = J - 1$ which implies $\ell = 1$ and formula (A.19) coincides with (A.15). Thus, we conclude that the characteristic function for the detailed coefficients for a fixed scale $l = J - j$ is

(6.3).

Q.E.D.