Wavelet Techniques for Time Series and Poisson Data

By

Piotr Z. Fryzlewicz

NISI DOMI QUIANUS

VIM·PROMOVET·INSITAM

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science

September 2003

Department of Mathematics

Abstract

This thesis considers the application of wavelet methods to a selection of problems arising in non-stationary time series analysis and Poisson regression.

In the first part of the thesis, we attempt to provide an answer to the question of whether and how wavelets can be useful in forecasting non-stationary time series. To achieve this, we consider several theoretical and computational aspects of forecasting in the Locally Stationary Wavelet (LSW) model (introduced by Nason *et al.* (2000)), which uses discrete non-decimated wavelets as building blocks. We propose a wavelet-based adaptive algorithm for forecasting non-stationary time series. The performance of the algorithm is investigated by simulation.

Secondly, we apply the LSW framework to model financial log-returns. We show that the LSW model accounts well for the stylised facts of log-return data. Several examples clearly demonstrate the need for local modelling of financial data, and also indicate the usefulness of wavelets as basic building blocks.

Next, we propose a multiscale algorithm for denoising the wavelet periodogram in the LSW model, and investigate some of its theoretical properties. The idea of the algorithm is to pre-process the data in the wavelet domain, in order to transform a gamma-contaminated signal into an approximately Gaussian-contaminated one, and then use one of the many denoising techniques available for Gaussian data. Then, the inverse transformation yields an estimate of the original signal.

Finally, as another application of the same methodology, we propose an algorithm for denoising Poisson-contaminated signals. We analyse some of its theoretical properties, and use simulation to demonstrate its excellent performance.

Acknowledgements

First of all, I would like to thank the Statistics Group and the Department of Mathematics at the University of Bristol for providing a perfect environment in which to pursue research. In particular, I would like to thank my adviser, Professor Guy Nason, for all his help, enthusiasm, encouragement, and extremely valuable feedback on my work. I am also grateful to my second adviser, Professor Bernard Silverman, not only for all his help, but also for holding an engaging seminar series on wavelets which I had the opportunity to attend. My research would not have been possible without my sponsors, the University of Bristol and "Universities UK", whose generous financial support is gratefully acknowledged.

Most of the second year of my study was spent collaborating on a joint research project with Professor Rainer von Sachs and Mr Sébastien Van Bellegem from the Institute of Statistics, University of Louvain-la-Neuve, Belgium. The collaboration not only taught me how to work in a research team, but also reminded me of how differently statistics can be understood and pursued on both sides of the Channel. I am also personally indebted to my coauthors for their hospitality during my research visits to Louvain-la-Neuve.

In the course of the last three years, I have had many interesting conversations with several other members of the mathematical and statistical community: among others, Anestis Antoniadis, Albert Cohen, Rainer Dahlhaus, Nils Hjort, Maarten Jansen, Thomas Mikosch, Byeong Uk Park, Theofanis Sapatinas, Catalin Starica and Kamila Żychaluk. Each of those discussions was special in its own way, but they were all highly enjoyable, stimulating, and friendly.

On a more personal note, I wish to thank all of my friends, first and foremost from Bristol, but also from Wrocław and elsewhere, for making the last three years such a wonderful time. I refrain from listing their names here for fear that this section might otherwise exceed the others in length!

Finally, I would like to thank my parents and sister for their constant love and support.

Dziękuję Wam wszystkim! Thank you all!

Declaration

I, the author, declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

The views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

Piotr Z. Fryzlewicz

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Wavelets can be casually described as oscillatory basis functions, cleverly constructed to possess several attractive features not enjoyed by "big waves" (sines and cosines): for example multiscale structure, ability to represent a variety of functions in a sparse manner, or simultaneous localisation in time and frequency. These and other properties have recently led many researchers to investigate the potential for using wavelets in various branches of statistics, such as time series analysis or nonparametric regression. In this thesis, we also employ wavelets to tackle a selection of problems arising in these two important areas of statistics.

In the introductory Chapter 2, we first review the basics of wavelet theory, and then provide a survey of wavelet applications in time series analysis and nonparametric regression. In particular, we describe the Locally Stationary Wavelet (LSW) time series model (Nason *et al.* (2000)), whose various aspects are studied in Chapters 3, 4 and 5. The LSW model uses wavelets as building blocks, which makes it a potentially useful tool for modelling multiscale phenomena whose characteristics evolve over time. Also, it uses the concept of rescaled time: the time-varying second order quantities are modelled as functions defined on a compact interval, which enables meaningful asymptotics. Chapter 2 concludes with a brief section on estimating Poisson intensities by wavelet methods, which prepares the ground for the material of Chapter 6.

The work of Chapter 3 is motivated by the interesting question of whether and how wavelets can help in forecasting non-stationary time series. We provide an answer by considering various aspects of linear prediction in the Gaussian LSW model. The rescaled time principle enables us to obtain a variety of asymptotic results. In particular, we generalise the Yule-Walker equations (well known in the stationary case), and derive Kolmogorov's formula for one-step prediction error. In the second half of Chapter 3, we analyse the properties of wavelet-based estimators of the prediction matrix, and provide a complete algorithm for forecasting non-stationary time series. Interesting and encouraging results are obtained by applying the algorithm to a meteorological time series.

In Chapter 4, we model financial log-return series as LSW processes. In our choice of model, we are motivated by several factors, including the comment made in Calvet & Fisher (2001) that various economic agents operate at different time scales, which may translate into a possible multiscale mechanism underlying financial log-returns. We slightly modify the definition of the LSW model to include the time-modulated white noise, the simplest possible linear model for log-returns, as a special case. We then exploit the rescaled time principle to provide a theoretical explanation of the "stylised facts" of financial time series in the LSW framework. We propose a generic algorithm for estimating the time-varying covariance structure of log-returns, and perform various analyses of log-return data in the LSW framework. These seem to confirm the appropriateness of the LSW model for the analysis of this type of data.

The work of Chapters 5 and 6 stems from a rather unexpected discovery that a computationally straightforward modification of the Discrete Haar Transform can be used to stabilise the variance of χ^2 and Poisson data. Being able to denoise the wavelet periodogram is essential for understanding the local second order structure of the LSW series under consideration. In Chapter 5, we propose a transformation, called the Haar-Fisz transform, for stabilising the variance of the wavelet periodogram in the Gaussian LSW model, and bringing its distribution closer to normality. Several theoretical results are established, and the above properties of the Haar-Fisz transform are proved under a certain asymptotic regime. The Haar-Fisz transform is shown to perform excellently in practice. Then, a denoising methodology for the wavelet periodogram is proposed, which consists of taking the Haar-Fisz transform of the periodogram, denoising the transformed vector using any technique suitable for Gaussian data, and taking the inverse Haar-Fisz transform. Simulations and an example involving the Dow Jones series demonstrate the usefulness of the technique.

The Haar-Fisz transform constitutes a "bridge" between Chapters 5 and 6. In the latter, we propose a similar technique for stabilising the variance of sequences of Poisson counts and bringing their distribution close to Gaussianity. The Haar-Fisz transform for Poisson data is investigated theoretically for constant intensities, and empirically for non-constant ones. It turns out that in this context, the Haar-Fisz transform is a more effective Gaussianiser and variance stabiliser than the traditional square-root transform. A Haar-Fisz-based algorithm for Poisson intensity estimation is proposed: its performance is shown to be typically better than, but occasionally comparable to, that of the current state-of-the-art techniques.

Finally, Chapter 7 concludes with a summary of contributions and a few interesting ideas for future research.

Chapter 2

Literature review

This chapter provides an overview of the wavelet theory and reviews recent literature in the two areas of statistics studied in this thesis: time series analysis and nonparametric regression. We place particular emphasis on evolutionary spectral theory for time series and multiscale methods for Poisson data.

2.1 Wavelets

Wavelets can be informally described as localised, oscillatory functions designed to have several attractive properties not enjoyed by "big waves" — sines and cosines. Since their "invention" in the early eighties (the term "wavelet" appeared for the first time in Morlet *et al.* (1982)), wavelets have received enormous attention both in the mathematical community and in the applied sciences. Several monographs appeared, both on the mathematical theory of wavelets (Meyer (1992), Daubechies (1992), Chui (1992), Mallat (1998), Cohen (2003)), as well as on their applications (Ruskai (1992), Jaffard *et al.* (2001)).

Formally, after Daubechies (1992), we define a *wavelet* to be any function $\psi \in \mathbb{L}_2(\mathbb{R})$ which satisfies the admissibility condition

$$\int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty. \tag{2.1}$$

In this thesis, we only concentrate on those wavelet functions whose dyadic dilations and translations

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z} \tag{2.2}$$

form an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$. Indices j and k are commonly called scale (or dilation) and location (or translation) parameters, respectively. Condition (2.1) implies, in particular, that

$$\int_{\mathbb{R}} \psi(x)dx = 0. \tag{2.3}$$

While (2.1) can be viewed as a requirement that ψ should be localised in frequency, (2.3) can be interpreted as both localisation in time (as it implies $\psi \in \mathbb{L}_1(\mathbb{R})$) and oscillation.

Haar (1910) considered an orthonormal basis which would later become arguably the best-known wavelet system with the wavelet function of the form

$$\psi^H(x) = \mathbb{I}_{\{0 \le x < 1/2\}} - \mathbb{I}_{\{1/2 \le x \le 1\}}. \tag{2.4}$$

We say that the wavelet ψ has n vanishing moments if

$$\int_{-\infty}^{\infty} x^k \psi(x)dx = 0 \quad \text{for} \quad k \in \{0, 1, \dots, n\}. \tag{2.5}$$

It is easy to see that ψ^H has 0 vanishing moments. Daubechies (1992, Chapter 6) identifies the *Extremal Phase* family: a collection of orthonormal wavelet bases possessing different degrees of smoothness and numbers of vanishing moments. This family of bases is indexed by the number of vanishing moments and the Haar basis is its zeroth member. A review of this and other families of wavelets (including Daubechies' *Least Asymmetric* family) can be found in Vidakovic (1999), Sections 3.4 and 3.5.

The vanishing moments property (2.5), together with the localisation properties (2.1) and (2.3), imply that wavelets are often capable of representing signals in a sparse manner. Coefficients $d_{j,k}$ of the wavelet expansion of $f \in \mathbb{L}_2(\mathbb{R})$ can be expressed in the usual way as $d_{j,k} = \langle f, \psi_{j,k} \rangle$. Larger (smaller) values of j correspond to *finer (coarser) scale* coefficients.

2.1.1 Multiresolution analysis

In statistics, we are often faced with discretely-sampled signals and therefore we need to be able to perform wavelet decomposition of vectors, rather than continuous functions as above. The *multiresolution analysis* framework, first introduced by Mallat (1989a,b), is commonly used to define discrete wavelet filters. The starting point is a scaling function ϕ and a multiresolution analysis of $\mathbb{L}_2(\mathbb{R})$, i.e. a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $\mathbb{L}_2(\mathbb{R})$ such that

- $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0;
- $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots \subset \mathbb{L}_2(\mathbb{R})$;
- $f \in V_j \iff f(2\cdot) \in V_{j+1}$;
- $\bigcap_j V_j = \{0\}$, $\overline{\bigcup_j V_j} = \mathbb{L}_2(\mathbb{R})$.

The set $\{\sqrt{2}\phi(2x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_1 since the map $f \mapsto \sqrt{2}f(2\cdot)$ is an isometry from V_0 onto V_1. The function ϕ is in V_1 so it must have an expansion

$$\phi(x) = \sqrt{2}\sum_k h_k \phi(2x-k), \quad \{h_k\}_k \in l_2, \quad x \in \mathbb{R}. \tag{2.6}$$

Once we have the scaling function ϕ, we use it to define the wavelet function (also called the *mother wavelet*) ψ. We define the latter in such a way that $\{\psi(x-k)\}_k$ is an orthonormal basis for the space W_0, being the orthogonal complement of V_0 in V_1:

$$V_1 = V_0 \oplus W_0. \tag{2.7}$$

Defining $W_j = \mathrm{span}\{\psi_{j,k} : k \in \mathbb{Z}\}$, we obtain that W_j is the orthogonal complement of V_j in V_{j+1}. We can write

$$V_{j+1} = V_j \oplus W_j = \ldots = V_0 \oplus \left(\bigoplus_{i=0}^{j} W_i\right), \tag{2.8}$$

or, taking the limit (recall that $\bigcup_j V_j$ is dense in $\mathbb{L}_2(\mathbb{R})$),

$$\mathbb{L}_2(\mathbb{R}) = V_0 \oplus \left(\bigoplus_{i=0}^{\infty} W_i \right) = V_{j_0} \oplus \left(\bigoplus_{i=j_0}^{\infty} W_i \right), \quad \forall j_0. \tag{2.9}$$

There are precise procedures for finding ψ once ϕ is known (see Daubechies (1992), Section 5.1). One possibility (Daubechies (1992), Theorem 5.1.1) is to set

$$\psi(x) = \sqrt{2} \sum_k h_{1-k} (-1)^k \phi(2x - k). \tag{2.10}$$

It can be shown that the appropriate orthogonality conditions are satisfied.

2.1.2 Discrete Wavelet Transform

The nested structure of the multiresolution analysis can be exploited to construct a fast decomposition-reconstruction algorithm for discrete data, analogous to the Fast Fourier Transform of Cooley & Tukey (1965). The algorithm, called the *Discrete Wavelet Transform* (Mallat (1989a,b)) produces a vector of wavelet coefficients of the input vector at dyadic scales and locations. The transformation is linear and orthonormal but is not performed by matrix multiplication to save time and memory.

We first describe a single *reconstruction* step, used in computing the inverse Discrete Wavelet Transform (DWT). The following two sets are orthonormal bases for V_1: $\{\sqrt{2}\phi(2x - k)\}_{k \in \mathbb{Z}}$, $\{\phi(x - k), \psi(x - l)\}_{k,l \in \mathbb{Z}}$. Using (2.6) and (2.10), we obtain for any $f \in V_1$

$$\begin{aligned} f(x) &= \sum_k c_{0,k} \phi(x - k) + \sum_k d_{0,k} \psi(x - k) \\ &= \sum_l \left(\sum_k h_l c_{0,k} + \sum_k h_{1-l} (-1)^l d_{0,k} \right) \sqrt{2} \phi(2x - 2k - l) \\ &= \sum_{l'} \left(\sum_k h_{l'-2k} c_{0,k} + \sum_k h_{1-l'+2k} (-1)^{l'} d_{0,k} \right) \sqrt{2} \phi(2x - l'). \end{aligned}$$

Writing the expansion w.r.t. the other basis as $f(x) = \sum_{l'} c_{1,l'} \sqrt{2} \phi(2x - l')$ and equating the coefficients, we obtain

$$c_{1,l'} = \sum_k h_{l'-2k} c_{0,k} + \sum_k h_{1-l'+2k} (-1)^{l'} d_{0,k}, \tag{2.11}$$

which completes the reconstruction part: the coarser scale coefficients $\{c_{0,k}\}$, $\{d_{0,k}\}$ are used to obtain the finer scale coefficients $\{c_{1,k}\}$.

The *decomposition* step (used in the DWT) is equally straightforward: we have

$$\begin{aligned} c_{0,k} &= \int_{-\infty}^{\infty} f(x)\phi(x-k)dx \\ &= \int_{-\infty}^{\infty} f(x)\sum_l h_l\sqrt{2}\phi(2x-2k-l)dx \\ &= \sum_l h_l c_{1,2k+l} = \sum_l c_{1,l}h_{l-2k}. \end{aligned} \tag{2.12}$$

Similarly,

$$d_{0,k} = \sum_l (-1)^{l-2k} h_{1-l+2k} c_{1,l}. \tag{2.13}$$

The same mechanism works for each scale: $\{c_{j,k}\}$ gives $\{c_{j-1,k}\}$ and $\{d_{j-1,k}\}$ for all j. On the other hand, $\{c_{j,k}\}$ can be reconstructed using $\{c_{j-1,k}\}$ and $\{d_{j-1,k}\}$ for all j. To start this "pyramid" algorithm, we only need to compute the scaling coefficients $c_{j,k}$ at the finest scale of interest, say $j = J$. Indeed, when performing wavelet decomposition of finite sequences, it is commonly assumed that our input vector $\mathbf{f} = \{f_n\}_{n=0}^{2^J-1}$ is a vector of scaling coefficients of a function f, i.e. $f_n = c_{J,n} = \langle f, \phi_{J,n}\rangle$, where $\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$. The DWT of $\mathbf{f}$ is given by

$$\mathrm{DWT}(\mathbf{f}) = (c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, \ldots, d_{2,3}, \ldots, d_{J-1,0}, \ldots, d_{J-1,2^{J-1}-1}). \tag{2.14}$$

Informally speaking, the wavelet coefficients $d_{j,k}$ contain information on the local oscillatory behaviour of $\mathbf{f}$ at scale j and location $2^{J-j}k$, whereas the coefficient $c_{0,0}$ contains information on the global "mean level" of $\mathbf{f}$. A few remarks are in order.

Decimation. Define

$$\begin{aligned} c^*_{0,k} &= \sum_l c_{1,l}h_{l-k} \\ d^*_{0,k} &= \sum_l (-1)^{l-k} h_{1-l+k} c_{1,l}, \end{aligned}$$

so that $c^*_{0,k}$ is a convolution of $c_{1,k}$ with h_k, and $d^*_{0,k}$ is a convolution of $c_{1,k}$ with $(-1)^k h_{1-k}$. We have $c_{0,k} = c^*_{0,2k}$ and $d_{0,k} = d^*_{0,2k}$: coarser scale coefficients are *decimated* convolutions of finer scale coefficients with fixed (scale-independent) filters. This is in contrast to the *Non-decimated Wavelet Transform* where no decimation is performed, yielding a shift-invariant (but redundant) transform: see Section 2.1.3 for details.

High-pass and low-pass filters. We define $g_k = (-1)^k h_{1-k}$. Due to its effect in the frequency domain, g_k (h_k) is often referred to as a *high-pass (low-pass) filter* in the wavelet literature (Daubechies (1992)). This motivates the commonly used name for the wavelet and scaling coefficients: they are often referred to as *detail* and *smooth* coefficients, respectively.

Example of the DWT. By simple algebra, $\phi^H(x) = \mathbb{I}_{\{0 \le x \le 1\}}$ generates the Haar wavelet ψ^H, with a low-pass filter h_k s.t. $h_0 = h_1 = 1/\sqrt{2}$, $h_k = 0$ otherwise, and a high-pass filter g_k s.t. $g_0 = -g_1 = 1/\sqrt{2}$, $g_k = 0$ otherwise. We shall now decompose a four-element vector

$$(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3}) = (1, 1, 2, 3)$$

using the DWT with Haar wavelets. By (2.12) and (2.13), we obtain

$$\begin{aligned}
c_{1,0} &= 1/\sqrt{2} \times 1 + 1/\sqrt{2} \times 1 = \sqrt{2} \\
c_{1,1} &= 1/\sqrt{2} \times 2 + 1/\sqrt{2} \times 3 = 5/\sqrt{2} \\
d_{1,0} &= 1/\sqrt{2} \times 1 - 1/\sqrt{2} \times 1 = 0 \\
d_{1,1} &= 1/\sqrt{2} \times 2 - 1/\sqrt{2} \times 3 = -1/\sqrt{2}.
\end{aligned}$$

Continuing at the next coarser scale, we obtain

$$\begin{aligned}
c_{0,0} &= 1/\sqrt{2} \times \sqrt{2} + 1/\sqrt{2} \times 5/\sqrt{2} = 7/2 \\
d_{0,0} &= 1/\sqrt{2} \times \sqrt{2} - 1/\sqrt{2} \times 5/\sqrt{2} = -3/2.
\end{aligned}$$

The original vector $(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3})$ can now be easily reconstructed from $(c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1})$, (i.e. from the smooth coefficient at the coarsest scale and the detail coefficients at all scales) using the inverse DWT. As the DWT is orthonormal, the inverse DWT uses exactly the same filters as the DWT.

Note that the high-pass filter annihilates constants (recall that Haar wavelets have vanishing moments up to degree 0). Wavelets with higher numbers of vanishing moments are capable of annihilating polynomials of higher degrees.

Boundary issue. With wavelet filters longer than Haar, there often arises the problem of what action to perform when the support of the filter extends beyond the support of the input vector. Several solutions have been proposed, including symmetric reflection of the input vector at the boundaries, polynomial extrapolation, periodising the vector, padding it out with zeros, etc. See Nason & Silverman (1994) for an overview. Cohen *et al.* (1993) introduced *wavelets on the interval*, i.e. wavelet bases for functions defined on an interval as opposed to the whole real line. They also proposed a corresponding fast wavelet transform which uses filters adapted to the finite support situation. The lifting scheme (see Section 2.1.4) offers a natural way of dealing with the boundary problem.

Computational speed. $O(n)$ operations are needed for the DWT which uses a compactly-supported wavelet, where n is the size of the input sequence. This is an advantage over the Fast Fourier Transform, which requires $O(n\log(n))$ operations.

2.1.3 Non-decimated Wavelet Transform

An undesirable property of the DWT is that it is not translation-invariant, and that at any given scale, it only provides information about the input vector at certain (dyadic) locations. Using the toy example above, the coefficient $c_{1,0}$ uses $c_{2,0}$ and $c_{2,1}$, while the coefficient $c_{1,1}$ uses $c_{2,2}$ and $c_{2,3}$, but there is no coefficient which would use, say, $c_{2,1}$ and $c_{2,2}$. Motivated by this, Pesquet *et al.* (1996) introduce a Non-decimated DWT (NDWT) which remedies this problem by computing wavelet coefficients at all possible locations at all scales (see also Nason & Silverman (1995), Coifman & Donoho (1995)). Continuing the example of the previous section, the NDWT of $(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3}) = (1, 1, 2, 3)$ which uses Haar wavelets is performed as follows. We begin with

$$\begin{aligned}
c_{1,0} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,0}, c_{2,1}) \\
c_{1,1} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,1}, c_{2,2}) \\
c_{1,2} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,2}, c_{2,3}) \\
c_{1,3} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,3}, c_{2,0}),
\end{aligned}$$

where the "$\cdot$" denotes the dot product. The detail coefficients $d_{1,k}$ are obtained similarly by replacing the low-pass filter with the high-pass one. Note that we implicitly assume "periodic" boundary conditions in the above (see the remark on the "boundary issue" in Section 2.1.2). Before we proceed to the next stage, we insert zeros between each two elements of the wavelet filters. Thus, we have

$$\begin{aligned}
c_{0,0} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,0}, c_{1,1}, c_{1,2}, c_{1,3}) \\
c_{0,1} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,1}, c_{1,2}, c_{1,3}, c_{1,0}) \\
c_{0,2} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,2}, c_{1,3}, c_{1,0}, c_{1,1}) \\
c_{0,3} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,3}, c_{1,0}, c_{1,1}, c_{1,2}),
\end{aligned}$$

and similarly for the detail coefficients. The insertion of zeros is necessary since decimation is not performed. Were we to compute the NDWT at yet another scale, we would use the filter $(1/\sqrt{2}, 0, 0, 0, 1/\sqrt{2}, 0, 0, 0)$ for the smooth and $(1/\sqrt{2}, 0, 0, 0, -1/\sqrt{2}, 0, 0, 0)$ for the detail. The computational speed of the NDWT is $O(n \log(n))$, where n is the length of the input vector.

2.1.4 Recent extensions of wavelets

Since the late eighties, several extensions and modifications of wavelets have been proposed. We only give a brief overview below.

The construction of *multidimensional wavelets* is due to Mallat (1989a), who also proposed a multivariate version of the DWT. Cohen *et al.* (1992) introduced *biorthogonal wavelets*, where the decomposition and reconstruction steps use different non-orthogonal bases which are however, in a certain sense, mutually orthogonal. Geronimo *et al.* (1994) formulated *multiple wavelets* which use translations and dilations of more than one wavelet function. Lawton (1993) derived *complex-valued wavelets* (although their construction was already mentioned in Daubechies (1992); see also Lina & Mayrand (1995) for a detailed description and derivation of complex Daubechies' wavelets). Coifman *et al.* (1989) introduced *wavelet packets*: redundant collections of linear combinations of wavelets capable of representing signals more economically than wavelets themselves. Wavelet packet coefficients are rapidly computable by applying both low- and high-pass filters to both smooth and detail coefficients, and can be searched for the "best basis" representation (Coifman & Wickerhauser (1992)).

Donoho (2000) introduced orthonormal *ridgelets*, which form a basis of $\mathbb{L}_2(\mathbb{R}^2)$ and provide an efficient representation of so-called ridge functions, i.e. functions of the form $r_\theta(x) = r(x_1 \cos(\theta) + x_2 \sin(\theta))$, where the ridge profile r is not necessarily smooth. *Curvelets* (Candès & Donoho (2001)), whose theory relies on ridgelets, provide a near-optimal approximation of distributions in 2D which are integrals along curves and which can be viewed as 2D extensions of the Dirac delta. Donoho & Huo (2002) introduced *beamlets*, i.e. collections of line segments in 2D, occurring at dyadic locations and scales and at a range of orientations; a beamlet transform of a 2D function is a collection of integrals along beamlets.

The *lifting scheme*, proposed by Sweldens (1996) is a powerful way of generating multiscale transforms of (possibly) unequally spaced data. The transform consists of predict and update steps, is fast and can be performed "in place". For particular choices of filters, the lifting scheme generalises the (bi)orthogonal DWT or the wavelet packet transform.

Some of the above extensions, as well as some others, are discussed in detail in

Vidakovic (1999), Chapter 5.

2.1.5 Applications of wavelets

Wavelets and their extensions have been applied in a multitude of areas, such as signal and image processing, data compression, computer graphics, astronomy, quantum mechanics and turbulence: for a discussion of these and other areas of application we refer the reader to the monographs by Ruskai (1992) and Jaffard *et al.* (2001). An important field of application is numerical analysis, extensively covered in Cohen (2003). One can venture to say that wavelets are indeed one of those fortunate mathematical concepts that have almost become "household objects": for example, they were used in the JPEG2000 compression algorithm; and multiscale subdivision schemes, related to wavelets, were employed in some recent animated movies such as "A Bug's Life" (Mackenzie (2001)). See Cohen (2003), Chapter 2, for an overview of subdivision schemes and related topics.

Following Vidakovic (1999), who gives a comprehensive overview of wavelet applications in statistics, we list some of the most important areas of statistics where wavelets have been successfully applied:

- time series analysis (see Section 2.2 for more details);

- non-parametric regression (see Section 2.3 for more details);

- estimation of densities (Hall & Patil (1995), Donoho *et al.* (1996), Penev & Dechevsky (1997), Pinheiro & Vidakovic (1997), Antoniadis *et al.* (1999), Pensky (1999), Herrick *et al.* (2001)) and density functionals (Kerkyacharian & Picard (1996), Prakasa Rao (1999));

- deconvolution and inverse problems (Donoho (1995), Abramovich & Silverman (1998), Pensky & Vidakovic (1999), Walter & Shen (1999), Pensky (2002));

- statistical turbulence (see Farge *et al.* (1999) and Schneider & Farge (2001) for reviews).

Abramovich *et al.* (2000) is a useful review article on statistical applications of wavelets. We give a detailed review of wavelet applications in time series analysis (with a particular emphasis on evolutionary spectral theory) and non-parametric regression in Sections 2.2 and 2.3, respectively.

2.1.6 Summary

In this section, we briefly summarise the attractive features of wavelets shown or mentioned in this chapter.

Due to their vanishing moments property and localisation in time, wavelets are capable of representing certain functions, e.g. piecewise polynomials, in a sparse manner. As the wavelet coefficients computed at locations where the function is smooth will be zero, only a few significant coefficients will suffice to accurately approximate the function. Also, their simultaneous localisation in time and frequency makes them potentially useful building blocks for phenomena whose spectral characteristics evolve over time.

The multiscale structure inherent to wavelets serves two useful purposes: it enables the construction of fast decomposition-reconstruction algorithms, and it makes them a natural tool for analysing multiscale phenomena. The fact that it is possible to construct orthonormal wavelet bases is extremely important in statistics, where i.i.d. Gaussian noise in the time domain gets mapped to noise with the same characteristics in the wavelet domain.

Also, in contrast to Fourier analysis where only one set of basis functions is available, there are several families of wavelets to choose from.

2.2 Time series analysis

2.2.1 Introduction

A *time series* is a collection of random variables $\{X_t, t \in D \subseteq \mathbb{Z}\}$, with t often interpreted as time. Usually, $D = \mathbb{N}$, $D = \mathbb{Z}$ or $D = \{1, 2, \ldots, T\}$. With a slight abuse of terminology, an observed realisation of X_t is also often referred to as a time series. Time series arise is several areas of science and technology and time series analysis (TSA) is one of the most widely studied branches of statistics, with the *Journal of Time Series Analysis* dedicated solely to this important field. Two recommendable monographs are Priestley (1981) and Brockwell & Davis (1987).

We say that a time series X_t is *stationary* when (some of) its statistical properties do not change through time, therefore enabling us, in most cases, to estimate its parameters consistently. We say that X_t is *strict-sense stationary* if

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{D}{=} (X_{t_1+d}, \ldots, X_{t_n+d})$$

for all n, $t_1, \ldots, t_n$ and d. Often, strict-sense stationarity is too difficult to verify and/or too restrictive; one of the weaker concepts is that of *covariance stationarity.* For a univariate, zero-mean time series X_t, we define its *covariance* as $\tilde{\gamma}_X(s,t) = \mathbb{E}(X_s \overline{X_t})$. We say that X_t is covariance stationary if $\tilde{\gamma}_X(s,t) = \gamma_X(|t-s|)$.

It is well known (see e.g. Brockwell & Davis (1987), Theorem 4.8.2) that every univariate, zero-mean, covariance stationary discrete-time process has the following Cramér representation:

$$X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) dZ(\omega), \quad t \in \mathbb{Z}, \tag{2.15}$$

where $A(\omega)$ is the amplitude and $Z(\omega)$ is a stochastic process with orthonormal increments, i.e. $\mathbb{E}(dZ(\omega_1)\overline{dZ(\omega_2)}) = d\omega_1 \delta_{\omega_1 = \omega_2}$, where δ is the Kronecker delta. The parameter ω can be interpreted as frequency: X_t is a weighted linear combination of Fourier exponentials oscillating at various frequencies. Correspondingly, under mild conditions (Brockwell & Davis (1987), Theorem 4.9.2), the covariance function of X_t can be expressed as

$$\gamma_X(\tau) = \int_{-\pi}^{\pi} f_X(\omega) \exp(i\omega\tau) d\omega, \tag{2.16}$$

where $f_X(\omega) := |A(\omega)|^2$ is called the *spectral density* of X_t.

Below we list two of the most commonly used time series models. All the definitions given below hold for univariate, zero-mean processes.

ARMA models. ARMA (Autoregressive Moving Average) processes are arguably the most popular time series models used in the applied sciences. An ARMA(p, q) process X_t is defined as

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}, \tag{2.17}$$

where ε_t is a sequence of independent or uncorrelated identically distributed r.v.s, often assumed Gaussian for tractability. Stationarity of an ARMA(p, q) process is guaranteed by the condition that the polynomial

$$\alpha(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p \tag{2.18}$$

has no roots in the closed unit disk, e.g. $X_t = 0.9X_{t-1} + \varepsilon_t$ is stationary but $X_t = X_{t-1} + \varepsilon_t$ is not. The spectral density of X_t is given by

$$f_X(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{\beta(e^{-i\omega})}{\alpha(e^{-i\omega})} \right|^2, \tag{2.19}$$

where $\beta(z) = 1 + \beta_1 z + \ldots + \beta_p z^p$ and $\sigma^2 = \mathrm{Var}(\varepsilon_t)$.

ARMA(p, q) is an example of a so-called *linear* time series model, where X_t and its innovations ε_t are related by a linear mapping.

GARCH models. Several authors have studied financial log-return series, i.e. time series of the form $X_t = \log(P_t/P_{t-1})$, where P_t is a share price, a stock index, or a currency exchange rate. It has been empirically observed that most financial log-returns display the following features: the sample mean is close to zero; the marginal distribution is heavy-tailed; the sample autocorrelations of X_t are mostly insignificant but those of $|X_t|^p$ decay only very slowly;

finally there are "bursts" of high volatility (standard deviation) among periods of low volatility. These "stylised facts" imply that financial log-returns cannot be modelled as stationary linear processes: to preserve stationarity, various *non-linear* models have been proposed. The Autoregressive Conditionally Heteroscedastic (ARCH) model was proposed by Engle (1982), and Generalised ARCH (GARCH), its most popular extension — independently by Bollerslev (1986) and Taylor (1986). The Stochastic Volatility (SV) model was suggested by Taylor (1986) as an alternative to ARCH-type modelling. The two families of models are by far the most widely used in practice and there is massive literature on both of them; Cox *et al.* (1996) and Maddala & Rao (1996) are two recommendable monographs.

The zero-mean GARCH(p,q) model is specified as

$$\begin{aligned} X_t &= \sigma_t Z_t, \quad t \in \mathbb{Z} \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \end{aligned} \tag{2.20}$$

where Z_t is symmetric i.i.d. with variance one and $\alpha_i, \beta_i \geq 0$. In other words, the current standard deviation is a linear deterministic function of the past squared returns and/or the past values of the variance. By contrast, in the SV framework, the current variance is modelled as a stochastic function of the past returns.

Strict-sense stationarity of a GARCH(p, q) process is guaranteed by the well known conditions that $\alpha_0 > 0$ and

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \tag{2.21}$$

see Bougerol & Picard (1992). Davidson (2003) reviews some recent extensions to the ARCH model, analyses their moment and memory properties, and proposes a new model.

2.2.2 Evolutionary spectral theory

Time series which cannot be modelled well as stationary processes arise in several fields, e.g. biomedical TSA (Nason *et al.* (2000)) or geophysics (Sakiyama (2002)). Also in finance, several authors agree that stationary nonlinear processes cannot account well for some empirical characteristics of log-return data: see e.g. Mikosch & Starica (2003), Kokoszka & Leipus (2000) (who look at change point detection in ARCH models) or Härdle *et al.* (2000) (who introduce a time-varying SV model and look at the adaptive estimation of its parameters).

In this section, we review some of those time series models which assume that the process under consideration can be "well approximated", in some sense, by a stationary model over a short stretch of time. An appropriate name for this concept would be "local stationarity"; however, this term has already been reserved for a subclass of processes possessing this characteristic so we avoid using it at this stage. We only restrict ourselves to linear models and refer the reader interested in non-stationary nonlinear models to the articles on financial time series listed above.

Piecewise stationarity, possibly the simplest departure from stationarity, was considered e.g. by Ombao *et al.* (2001a), who attempted to find "optimal" stretches of stationarity in the series in a data-driven way. Several other approaches assumed a smoother evolution of the second-order structure. Here, two subgroups can be distinguished:

- time-domain approaches, which allow the coefficients of a parametric model, e.g. AR, to vary slowly with time: Mélard & Herteleer-De Schutter (1989), Dahlhaus *et al.* (1999), Grillenzoni (2000);

- frequency-domain approaches, which control the evolution of frequency-dependent quantities over time: Priestley (1965), Battaglia (1979), Dahlhaus (1997), Mallat *et al.* (1998), Swift (2000), Ombao *et al.* (2002).

Dahlhaus (1996a) introduces an important concept of *rescaled time* into the analysis of non-stationary time series. In his class of Locally Stationary Fourier (LSF) processes, X is modelled as a triangular stochastic array $\{X_{t,T}\}_{t=1}^{T}$, $T = 1, 2, \ldots$, such that

$$X_{t,T} = \int_{-\pi}^{\pi} \exp(i\omega t) A^0_{t,T}(\omega) dZ(\omega), \tag{2.22}$$

and there exists a function $A : [0, 1] \times (-\pi, \pi] \mapsto \mathbb{C}$, continuous in the first argument, such that

$$\sup_{t,\omega} \left| A^0_{t,T}(\omega) - A\left(\frac{t}{T}, \omega\right) \right| \leq \frac{K}{T} \quad \forall T \tag{2.23}$$

(see the paper for a complete definition). The time-rescaling in (2.23) is reminiscent of nonparametric regression where the function of interest is also defined on a finite interval and possesses some degree of regularity, thus enabling asymptotic considerations of e.g. consistency of the estimation procedure. In Dahlhaus' approach, the longer the stretch of the series, the finer the grid t/T and therefore the more information is gathered about $A(u, \omega)$ and about the evolutionary spectral density defined as $f(u, \omega) := |A(u, \omega)|^2$. Kim (1998) provides various statistical analyses of financial and macroeconomic data in the LSF framework.

The approach of Nason *et al.* (2000), which also adopts the rescaled time concept, will be discussed in detail in Section 2.2.4.

2.2.3 Wavelets and time series

Wavelets, due to their attractive properties listed in Section 2.1.6, have been used extensively in TSA. Reviews of wavelet methods in time series forecasting and wavelet smoothing in TSA appear in separate sections (Section 2.2.5 and 2.3.2, respectively). The paper by Nason & von Sachs (1999) reviews the use of wavelets in TSA, and the comprehensive monograph of Percival & Walden (2000) covers, among others, wavelet analysis of long memory processes (see also Vidakovic (1999), Section 9.5).

Several authors use wavelets in hypothesis testing in TSA: for example, Neumann & von Sachs (2000) propose a test for time series stationarity, Lee & Hong (2001) construct a test for serial correlation, and Whitcher *et al.* (2002) propose a test for variance homogeneity in long memory processes. Chiann & Morettin (1999) define the *wavelet periodogram* for stationary processes as a sequence of squared wavelet coefficients of the process; they also analyse some of its properties. Even though this wavelet-based analysis provides useful insight into the data, it is the classical Fourier analysis that can be shown to be "optimal" for stationary processes, see e.g. Priestley (1965). Nason & Sapatinas (2002) use wavelet packets to model a transfer function between two nonstationary time series. Wang *et al.* (2001) and Audit *et al.* (2002), among others, use wavelets to estimate the scaling exponent in self-similar processes. Bilen & Huzurbazar (2002) propose a model-free method for detecting outliers in time series data using wavelets, Wong *et al.* (2001) use wavelets to detect jumps and Li & Xie (1997) — hidden periodicities. Walden & Serroukh (2002) construct multi-resolution filters for the analysis of matrix-valued time series. Whitcher (2001) proposes a method, based on wavelet packets, for simulating Gaussian processes with unbounded spectra. Serroukh *et al.* (2000) investigate time-scale properties of time series in various models by estimating the variance of non-decimated wavelet coefficients (so-called "wavelet variance") at different scales. Rao & Indukumar (1996) look at higher order moments of wavelet transforms of nonlinear signals.

In financial time series, Hong & Lee (2001) develop a test for ARCH effects using a wavelet estimator of the spectral density of the squared residuals at frequency zero. Struzik (2001) uses wavelets to examine the scaling properties of the S&P index, and Gencay *et al.* (2001) — those of foreign exchange volatility. Ramsey (1999) and Ramsey (2002) review the use of wavelet analysis in finance and economics. Gençay *et al.* (2001) is an introductory monograph on wavelet methods in finance and economics.

2.2.4 The Locally Stationary Wavelet model

We now move on to describe the time series model whose various aspects are studied in Chapters 3, 4 and 5 of this thesis. The Locally Stationary Wavelet (LSW) model, due to Nason *et al.* (2000), is based on two main ingredients:

- following Dahlhaus (1996a), it adopts the rescaled time principle;
- it replaces the representation with respect to the Fourier basis by a representation with respect to non-decimated discrete wavelets.

Before defining the LSW model, we first define compactly supported discrete wavelet vectors. In what follows, $j = -1$ denotes the finest scale, $j = -2$ is the second finest scale, etc. Following Nason *et al.* (2000), we define the *discrete wavelet vectors* associated with filters $\{h_k\}$, $\{g_k\}$ as $\psi_j = (\psi_{j,0}, \dots, \psi_{j,\mathcal{L}_j - 1})$, where

$$\begin{aligned} \psi_{-1,n} &= g_n \\ \psi_{j,n} &= \sum_k h_{n-2k}\psi_{j+1,k} \quad \text{for} \quad j < -1 \\ \mathcal{L}_j &= (2^{-j} - 1)(N_h - 1) + 1 \\ N_h &= \#\{k : h_k \neq 0\}. \end{aligned}$$

For example,

$$\psi^H_{j,n} = 2^{j/2}(\mathbb{I}_{\{0 \le n \le 2^{-j-1}-1\}} - \mathbb{I}_{2^{-j-1} \le n \le 2^{-j}-1}). \tag{2.24}$$

The nondecimated collection $\psi_{j,k}(t)$ of discrete wavelet vectors is formed by shifting vectors ψ_j to all integer locations k:

$$\psi_{j,k}(t) := \psi_{j,t-k}. \tag{2.25}$$

We are now in a position to define the LSW model.

Definition 2.2.1 (Nason *et al.* (2000)) *A triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$, for $T = 1, 2, \dots$, is in the class of LSW processes if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-J(T)}^{-1} \sum_{k=-\infty}^{\infty} \omega_{j,k;T}\psi_{j,k}(t)\xi_{j,k}, \tag{2.26}$$

where $\psi_{j,k}(t)$ *are nondecimated discrete wavelet vectors,* $\omega_{j,k;T}$ *are real constants,* $J(T) = -\min\{j : \mathcal{L}_j \leq T\}$*, and* $\{\xi_{j,k}\}_{j,k}$ *are zero-mean orthonormal identically distributed random variables. Also, we assume that for each* $j \leq -1$*, there exists a Lipschitz function* $W_j(z) : [0,1] \to \mathbb{R}$ *such that*

- $\sum_{j=-\infty}^{-1} |W_j|^2 < \infty$,
- *the Lipschitz constants* L_j *satisfy*

$$\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty, \tag{2.27}$$

- *there exists a sequence of constants* C_j *satisfying* $\sum_{j=-\infty}^{-1} C_j < \infty$ *such that, for each* T*,*

$$\sup_{k=0,\dots,T-1} |\omega_{j,k;T} - W_j(k/T)| \leq C_j/T \quad \text{for} \quad j = -1, \dots, -J(T). \tag{2.28}$$

The representation in (2.26) can be viewed as a "wavelet counterpart" of the classical Cramér representation (2.15). As wavelets are parametrised by scale j and location k, the integration over frequencies in (2.15) is replaced by the summation over j and k in (2.26). Thus, the representation becomes naturally location-dependent (or, in this case, time-dependent).

Like in the classical theory, $\omega^2_{j,k;T}$ (the square of the amplitude, or the transfer function) constitutes a "wavelet spectrum" which measures the power of the series at scale j and location k. Our aim will often be to make inference on this quantity; however, if $\omega^2_{j,k;T}$ depends on j, k in an arbitrary fashion, there is no hope of estimating it accurately: note that we only observe a single row of the triangular array $X_{t,T}$, and there are of order $J(T) \times T = O(T \log(T))$ parameters $\omega^2_{j,k;T}$ to be estimated. Clearly, we cannot do a good job here unless we control the evolution of this sequence, and this is where the rescaled time property (2.28) comes into play. It ensures that for each j, the sequence $\{\omega_{j,k;T}\}_{k=0}^{T-1}$ evolves slowly, by requiring that it should be "close" to a sequence formed by sampling a regular (here, Lipschitz continuous) function $W_j(z)$ on a finer and finer grid. This idea, adopted from Dahlhaus (1996a), embeds inference in the LSW framework into the nonparametric regression setting. Note that, unlike the classical setting, $T \to \infty$

does not mean that more and more future observations arrive; the rows of $\{X\}_{t,T}$ are completely different stochastic processes, only linked to each other by the fact that they possess the same asymptotic transfer function $W_j(z)$.

The asymptotic *evolutionary wavelet spectrum* $S_j(z)$ is defined in rescaled time as

$$S_j(z) = W_j(z)^2 = \lim_{T\to\infty} \omega^2_{j,\lfloor zT \rfloor;T}. \tag{2.29}$$

In the classical theory, the spectral density and the covariance function are Fourier transforms of each other, see formula (2.16). It is possible to establish an analogous relationship here. Let $c_T(z,\tau)$ denote the finite-sample covariance function of $X_{t,T}$ at lag τ and rescaled time location z:

$$c_T(z,\tau) = \mathbb{E}(X_{\lfloor zT \rfloor,T} X_{\lfloor zT \rfloor+\tau,T}). \tag{2.30}$$

Further, let us recall the definition of *autocorrelation wavelets* Ψ_j from Nason *et al.* (2000):

$$\Psi_j(\tau) = \sum_k \psi_{j,k}\psi_{j,k+\tau}. \tag{2.31}$$

The system $\{\Psi_j\}_j$ is linearly independent, see Nason *et al.* (2000), Theorem 2.13. Some other properties of $\{\Psi_j\}_j$ will be discussed in Chapter 3. Let $c(z,\tau)$ denote the *asymptotic local covariance* function of $X_{t,T}$ at lag τ and rescaled time location z, defined as a transform of $\{S_j(z)\}_j$ with respect to the set of autocorrelation wavelets:

$$c(z,\tau) = \sum_{j=-1}^{-\infty} S_j(z)\Psi_j(\tau). \tag{2.32}$$

We quote the following result.

Theorem 2.2.1 (Nason *et al.* (2000)) *Under the assumptions of Definition 2.2.1,* $\|c - c_T\|_{L_\infty} = O(T^{-1})$.

Therefore, the asymptotic local covariance c is a good approximation to the finite-sample covariance c_T. Formula (2.32) provides a multiscale decomposition of the covariance structure of $X_{t,T}$. As $\Psi_j(0) = 1$ for all j, the local variance decomposes as

$$\sigma^2(z) := c(z,0) = \sum_j S_j(z). \tag{2.33}$$

Also, the representation (2.32) is invertible: denoting

$$A_{i,j} = \sum_\tau \Psi_i(\tau)\Psi_j(\tau), \tag{2.34}$$

we obtain

$$S_j(z) = \sum_\tau \left(\sum_i \Psi_i(\tau) A_{i,j}^{-1} \right) c(z,\tau) \tag{2.35}$$

(see Nason *et al.* (2000), Theorem 2.15, for the proof of invertibility of A).

Proposition 2.17 of Nason *et al.* (2000) states that all stationary processes with absolutely summable covariance are LSW processes; for them, the spectrum S_j does not depend on the rescaled time z.

One way of performing inference on time-varying second-order quantities in the LSW framework is by using the *wavelet periodogram*, defined below.

Definition 2.2.2 (Nason *et al.* (2000)) *Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. The triangular stochastic array*

$$I_{t,T}^{(j)} = \left| \sum_s X_{s,T} \psi_{j,t-s} \right|^2 \tag{2.36}$$

is called the wavelet periodogram of $X_{t,T}$ at scale j.

In practice, the wavelet periodogram is not computed separately for each scale j but instead, we compute the full NDWT transform of the observed row of $X_{t,T}$ with periodic boundary conditions, and then square the wavelet coefficients to obtain $I_{t,T}^{(j)}$ for $t = 0, \ldots, T-1$ and $j = -1, -2, \ldots, -J(T)$.

We quote the following result:

Proposition 2.2.1 (Nason *et al.* (2000)) *We have*

$$\mathbb{E} I_{t,T}^{(j)} = \sum_{i=-\infty}^{-1} S_i\left(\frac{t}{T}\right) A_{i,j} + O(2^{-j}/T). \tag{2.37}$$

If, in addition, $X_{t,T}$ is Gaussian, then

$$\mathrm{Var}\left(I_{t,T}^{(j)}\right) = 2\left(\sum_{i=-\infty}^{-1} S_i\left(\frac{t}{T}\right) A_{i,j}\right)^2 + O(2^{-j}/T). \tag{2.38}$$

Formulas (2.37) and (2.38) imply that the wavelet periodogram is an inconsistent but asymptotically unbiased estimator of a quantity which is a linear transform of the wavelet spectrum. By (2.37), an estimate of $S_j(z)$ can be obtained by setting $\hat{S}_j(z) = \sum_{i=-1}^{-J(T)} I_{\lfloor zT \rfloor,T}^{(i)} A_{i,j}^{-1}$. Some properties of this estimator are analysed in Nason *et al.* (2000).

Figure 2.1 shows an example of an LSW process whose spectrum is only non-zero at scales -1 and -3. $S_{-1}(z)$ and $S_{-3}(z)$ (bottom plot) are chosen in such a way that the local variance $c(z, 0) = \sigma^2(z)$ is independent of z, but $c(z, \tau)$ varies with z for $\tau = 1, 2, \ldots, 7$. The top plot shows a sample path of length 512 simulated from this spectrum using Haar wavelets and Gaussian innovations. It is visibly non-stationary: the series oscillates more rapidly over the time intervals where the finer-scale spectrum $S_{-1}(z)$ dominates.

In the original paper by Nason *et al.* (2000), the authors apply the LSW model to a biomedical time series. In Chapter 4 of this thesis, we demonstrate the usefulness of LSW modelling by considering various analyses of financial log-return data in the LSW framework.

2.2.5 Forecasting

Forecasting the future behaviour of time series is, along with understanding the data generating mechanism, one of the main aims of TSA, and two journals: *Journal of Forecasting* and *International Journal of Forecasting* publish articles devoted exclusively to this important area. Having observed $X_1, \ldots, X_t$, the quantity of interest to the analyst is often the predictor $\hat{X}_{t+h}$ of X_{t+h} ($h > 0$) which minimises the Mean-Square Prediction Error (MSPE):

$$\mathrm{MSPE}(\hat{X}_{t+h}, X_{t+h}) = \mathbb{E}(\hat{X}_{t+h} - X_{t+h})^2. \tag{2.39}$$

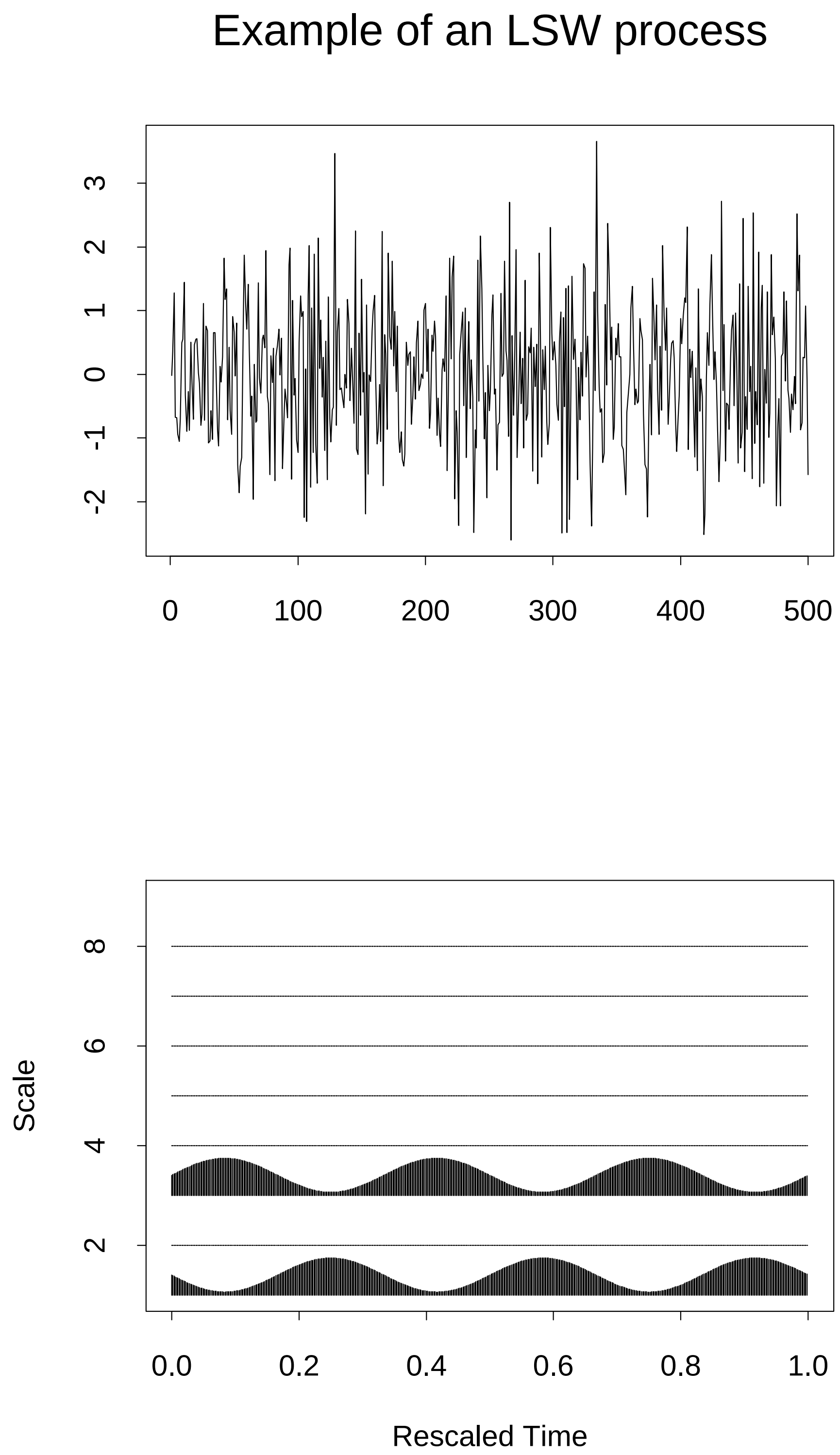

Figure 2.1: Bottom plot: spectrum of an exemplary LSW process plotted against the rescaled time. The y-axis shows negative scale $-j$. The spectrum is only non-zero at scales -1 and -3. Top plot: a sample path of length 512 simulated from this spectrum using Haar wavelets and Gaussian innovations.

(2.39) is minimised by

$$\hat{X}_{t+h} = \mathbb{E}(X_{t+h}|X_1, \ldots, X_t) \tag{2.40}$$

(see Brockwell & Davis (1987), Section 2.7). For Gaussian time series, $\hat{X}_{t+h}$ can be expressed as a linear combination of the past observations:

$$\hat{X}_{t+h} = \sum_{i=1}^{t} a_i X_i, \tag{2.41}$$

where a_i solve the so-called Yule-Walker equations

$$\mathrm{Cov}(X_n, X_{t+h}) = \sum_{i=1}^{t} a_i \,\mathrm{Cov}(X_n, X_i), \quad n = 1, \ldots, t, \tag{2.42}$$

which take a particularly simple form when X_t is stationary. GARCH models are used to forecast future volatility σ_{t+h} and not X_{t+h} itself: note that in the GARCH model specified by (2.20), the best mean-square predictor of X_{t+h}, given by (2.40), is simply zero. See Bera & Higgins (1993) for a discussion on forecasting in ARCH-type models.

For non-stationary Gaussian models, various more sophisticated forecasting techniques have been developed. *Kalman filtering* (see e.g. Chatfield (1996), Chapter 10) updates the parameters of the model as new observations arrive and can be used to produce forecasts. As well as being computationally fast, it exhibits fast convergence when the underlying model is stationary but is also able to trace the evolution of non-stationary models. *Bayesian forecasting* (West & Harrison (1997)) also exploits the principle of "parameter updating". Methods based on *neural networks* are often applied to the forecasting of non-linear time series, especially in the engineering literature, see e.g. Zhang *et al.* (2001b). Several other methods exist: the recent monograph by Chatfield (2000) provides a comprehensive overview.

Wavelets have often been used in time series forecasting in conjunction with neural network methods (Geva (1998), Milidiu *et al.* (1999), Hee *et al.* (2002), Soltani (2002)). Combined wavelet + neural network techniques were used to forecast electricity demand data (Zhang & Dong (2001)), financial time series (Zhang *et al.* (2001a)) and web traffic (Aussem & Murtagh (2001)). Some authors have considered forecasting based on wavelet methods but not supplemented with neural networks. The forecasting method proposed by Wong *et al.* (2003) relies on the decomposition of the time series using wavelets into three summands: trend, harmonic and irregular components. Li & Hinich (2002) use wavelets (and other filter banks) to forecast seasonal patterns. Zheng *et al.* (2001) apply their SVH-ARMA (state-dependent vector hybrid ARMA) technique to the forecasting of vector time series constructed by taking the DWT of scalar time series. Masuda & Okabe (2001) base their forecasting technique on the multiscale decomposition of a time series. The method of Soltani *et al.* (2000) exploits the decorrelating property of wavelets to forecast long-memory processes. Zheng *et al.* (2000) combine wavelets and Kalman filtering by modelling wavelet coefficients as state variables for the Kalman filter. Ikeda & Tokinaga (1999) use wavelets to forecast fractal time series.

In Chapter 3 of this thesis, we consider several theoretical and practical aspects of forecasting LSW processes reviewed in Section 2.2.4 above.

2.3 Nonparametric regression

In this section, we consider the problem of estimating a function $f : [0,1] \mapsto \mathbb{R}$ from noisy observations y_i on an equispaced grid:

$$y_i = f(i/n) + \epsilon_i, \quad i = 1, \dots, n, \tag{2.43}$$

where the ϵ_i's ("noise") are r.v.'s with $\mathbb{E}(y_i) = 0$. Denoting the estimator by $\hat{f} : [0,1] \mapsto \mathbb{R}$, we are often only interested in the values of $\hat{f}$ on $\{i/n\}_{i=1}^n$. The performance of $\hat{f}$ is often measured by the Mean-Square Error (MSE):

$$\text{MSE}(f, \hat{f}) = \frac{1}{n}\mathbb{E}\|\hat{f} - f\|_{l_2}^2. \tag{2.44}$$

Various subclasses of the problem can be identified, depending on the joint distribution of $\{\epsilon_i\}_{i=1}^n$ and on the smoothness of f. Linear methods produce an estimate $\hat{f}(i/n)$ by taking a linear transform of the observations: $\hat{f}(i/n) = B\mathbf{y}$, where B is a square matrix, and $\mathbf{y} = (y_1, \ldots, y_n)'$. Linear methods can often be shown to be optimal in terms of MSE if the underlying function f is smooth. For example, a linear method based on natural cubic splines is optimal for twice-differentiable functions in the sense that the estimator minimises the penalised sum of squares

$$S(f) = \sum_{i=1}^{n}(y_i - f(i/n))^2 + \alpha \int_0^1 (f''(x))^2 dx, \tag{2.45}$$

where the penalty term controls the "roughness" of f (see Green & Silverman (1994)). For reviews of other nonparametric linear methods, including kernel smoothing, see the monographs of Simonoff (1996) and Wand & Jones (1994).

2.3.1 Non-linear wavelet smoothing

For less regular (e.g. discontinuous) functions, linear smoothing performs inadequately, and non-linear smoothing methods are needed. In a seminal paper, Donoho & Johnstone (1994) introduce the principle of a non-linear smoothing method called *wavelet thresholding*. First, the signal is transformed via the DWT to obtain $d_{j,k} = \theta_{j,k} + \epsilon_{j,k}$, where $d_{j,k}$, $(\theta_{j,k}, \epsilon_{j,k})$ is the DWT of y_i $(f(i/n), \epsilon_i)$. Then, $d_{j,k}$ are shrunk towards zero (with the threshold chosen in an appropriate manner), and finally the inverse DWT is taken to obtain an estimate of f. The rationale behind this principle is twofold:

- As DWT is orthonormal, i.i.d. Gaussian noise in the time domain transforms into i.i.d. Gaussian noise in the wavelet domain;

- Due to the vanishing moments property, wavelet coefficients $\theta_{j,k}$ corresponding to the locations where the signal is smooth will be close to zero. On the other hand, those (hopefully few) corresponding to discontinuities or other irregularities will be significantly different from zero: the signal will be represented *sparsely* in the wavelet domain. Therefore, we can expect that an appropriately chosen threshold will be able to accurately separate signal from noise.

Two thresholding rules have been particularly commonly used and well-studied. For a given threshold λ, *hard* and *soft* thresholding shrink $d_{j,k}$ to

$$\begin{aligned} d^h_{j,k} &= d_{j,k}\mathbb{I}_{\{|d_{j,k}|>\lambda\}} \\ d^s_{j,k} &= \mathrm{sgn}(d_{j,k})(|d_{j,k}|-\lambda)_+, \end{aligned}$$

respectively. The threshold introduced in Donoho & Johnstone (1994) was the so-called *universal threshold*, $\lambda = \sigma\sqrt{2\log(n)}$. The authors show that the MSE of the soft thresholding estimator with the universal threshold is close (within a logarithmic factor) to the ideal risk one can achieve by "keeping" or "killing" the wavelet coefficients $d_{j,k}$ using knowledge of the underlying signal. At the same time, the universal threshold is an efficient noise suppressor as described in Section 4.2 of their paper.

In another ground-breaking paper, Donoho & Johnstone (1995) consider a non-linear wavelet estimator with soft thresholding where the threshold selection procedure is based on Stein's shrinkage method for estimating the mean of multivariate normal variables. They consider the behaviour of the estimator over a range of so-called Besov spaces (see Triebel (1983)), which form an extremely rich collection of functions with various degrees of smoothness (for certain values of the space parameters, Besov spaces can be shown to contain other better known function spaces such as Hölder or Sobolev spaces or the space of functions with bounded variation). The authors demonstrate that their estimator is *simultaneously nearly minimax* over a range of Besov balls, i.e. without knowing the regularity of the function, it nearly achieves the optimal rate of convergence which could be achieved if the regularity was known.

In most papers on the theory of non-linear wavelet estimation, it is assumed that the standard deviation σ of the noise is known. In practice, it needs to be estimated from the data. For Gaussian data, the method recommended by several authors (see e.g. Johnstone & Silverman (1997)) computes the scaled Median Absolute Deviation (MAD) on the sequence of wavelet coefficients at the finest resolution level, thereby ensuring robustness.

More recently, other thresholding rules have been proposed. Nason (1996) uses cross-validation as a means of selecting the threshold. Abramovich & Benjamini (1996) set up wavelet thresholding as a multiple hypothesis testing problem and propose an approach based on the so-called *false discovery rate*. Ogden & Parzen (1996) also adopt the hypothesis testing point of view and use recursive likelihood ratio tests to determine the threshold. Johnstone & Silverman (1997) consider level-dependent universal thresholding for correlated Gaussian noise. Averkamp & Houdré (2003) extend the approach of Donoho & Johnstone (1994) to other noise distributions such as exponential, mixture of normals or compactly supported distributions. Vanreas *et al.* (2002) consider stable wavelet transforms for denoising data observed on non-equispaced grids. Barber & Nason (2003) develop various thresholding procedures using complex-valued wavelets. Johnstone & Silverman (2003) propose an empirical Bayes approach to the threshold selection problem. Cai & Silverman (2001), among others, consider *block thresholding*: they propose a thresholding procedure whereby wavelet coefficients are considered in overlapping blocks and the action performed on the coefficients in the middle of the block depends upon the data in the whole block.

Coifman & Donoho (1995) introduce *translation invariant denoising*: the full NDWT transform of the data is taken, then the universal threshold is applied to all resulting wavelet coefficients, and then an inverse NDWT transform yields an estimate of the signal. As the NDWT is redundant, there are many possible ways of generating an inverse NDWT transform: the one proposed by the authors is equivalent to taking the average over all possible DWT's contained in the NDWT, corresponding to all possible circular shifts of the data set (hence the name "translation invariant").

2.3.2 Wavelet shrinkage in time series analysis

Wavelet shrinkage has been used extensively in the time series context. Gao (1997) proposes an algorithm for wavelet smoothing of the log-periodogram, using asymptotic normality of the wavelet coefficients at coarser scales, and adjusting the thresholds for non-normality at finer scales (in this case, the noise is asymptotically independent but not Gaussian). Neumann (1996) considers wavelet smoothing of a "tapered" periodogram for possibly non-Gaussian stationary time series, basing his choice of thresholds on asymptotic normality arguments. Neumann & von Sachs (1997) and von Sachs & Schneider (1996) propose thresholds for estimating time-varying spectrum in Dahlhaus' locally stationary time series model. von Sachs & MacGibbon (2000) consider wavelet thresholding of signals contaminated with locally stationary noise. Nason *et al.* (2000) propose the following threshold for shrinking the wavelet coefficients $\tilde{d}_{i,k}$ of the wavelet periodogram $I_{t,T}^{(j)}$ of a Gaussian LSW process:

$$\lambda_{i,k,j,T} = \sqrt{\mathrm{Var}(\tilde{d}_{i,k}) \log(T)}, \tag{2.46}$$

where a pre-estimate of each $\mathrm{Var}(\tilde{d}_{i,k})$ is required, which can potentially hamper the practical performance of the method. Also note that the threshold is independent of j. Cristan & Walden (2002) consider wavelet and wavelet packet smoothing of the (tapered and logged) periodogram, and conclude that the wavelet-based algorithm performs adequately and therefore the use of wavelet packets is not necessary (this article complements an earlier paper by Walden *et al.* (1998)). Truong & Patil (2001) derive MSE's of wavelet-based estimators of density and autoregression functions in stationary time series which satisfy appropriate mixing conditions. Dahlhaus & Neumann (2001) use wavelet shrinkage to estimate a time-varying p-dimensional parameter of the spectral density function of a non-stationary process. Hoffmann (1999) proposes a wavelet thresholding estimator of the mean and conditional variance functions in a non-linear AR(1) model.

In Chapter 5 of this thesis, we propose a multiscale technique for denoising the wavelet periodogram of a Gaussian LSW process.

2.3.3 Wavelet and multiscale methods for Poisson data

Some authors have also considered the problem of estimating the intensity of a Poisson process using a wavelet-based technique. The usual setting is as follows: the possibly inhomogeneous one-dimensional Poisson process is observed on the interval $[0, T)$, and discretised into a vector $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$, where v_n is the number of events falling into the interval $[nT/N, (n+1)T/N)$, and $N = 2^J$ is an integer power of two. Each v_n can be thought of as coming from a Poisson distribution with an unknown parameter λ_n, which needs to be estimated. Note that in this case the "noise" $v_n - \mathbb{E}(v_n)$ is independent but not identically distributed. The approach proposed by Donoho (1993) consists in first preprocessing the data using Anscombe's (1948) square-root transformation, $\mathcal{A}\mathbf{v} = 2\sqrt{\mathbf{v} + 3/8}$, so that the noise becomes approximately Gaussian. Then the analysis proceeds as if the noise were indeed Gaussian, yielding (after applying the inverse square-root transformation) an estimate of the intensity of the process.

Besbeas *et al.* (2004) report that the best performing methods currently available in literature are those based on translation-invariant multiscale Bayesian techniques as described in Kolaczyk (1999a) and Timmermann & Nowak (1997, 1999). Kolaczyk (1999a) introduces a Bayesian multiscale algorithm to estimate the discretised intensity. However, rather than transforming the data using a wavelet transform, he considers recursive dyadic partitions, and places prior distributions at the nodes of the binary trees associated with these partitions. The Bayesian methods outperform earlier techniques in Kolaczyk (1997, 1999b), Nowak & Baraniuk (1999) and also the recent technique of Antoniadis & Sapatinas (2001) (since the latter is equivalent to Nowak & Baraniuk (1999) for Poisson data). The article by Sardy *et al.* (2004) describes a computationally intensive l_1-penalised likelihood method which can be used for estimating Poisson intensities.

Other recent contributions to the field of wavelet-based intensity estimation include Patil & Wood (2004), who concentrate on the theoretical MSE properties of wavelet intensity estimators, where the intensity is a random process rather than a deterministic function (or, after discretisation, a deterministic vector). Brillinger (1998) gives a brief overview of wavelet-based methodology in the analysis of point process data, and obtains an estimate of the autointensity function of the well-known California earthquake data.

In Chapter 6 of this thesis, we propose a multiscale method for estimating the discretised intensity function of an inhomogeneous one-dimensional Poisson process.

Chapter 3

Forecasting LSW processes

In this chapter, we consider several theoretical and practical aspects of forecasting Gaussian LSW processes. Some results of this chapter were used, in a modified form, in the article by P. Fryzlewicz, S. Van Bellegem and R. von Sachs (2003) "Forecasting non-stationary time series by wavelet process modelling" (*Annals of the Institute of Statistical Mathematics*, **55**, 737–764). Throughout the thesis, this article will be referred to as Fryzlewicz *et al.* (2003). The results of the article which are not due to the author are only *quoted* in this chapter (without proofs) and their authorship is acknowledged.

The chapter is organised as follows. In Section 3.1, we investigate the minimisation of the approximate Mean Square Prediction Error (MSPE) for a linear predictor in the LSW framework. The reason why approximate MSPE minimisation is preferred is that it involves the uniquely defined asymptotic wavelet spectrum $\{S_j(z)\}_j$, unlike the exact MSPE which involves the unidentifiable finite-sample parameters $\omega_{j,k;T}$. In Section 3.2, we look in detail at the assumptions made in deriving the results of Section 3.1. We identify an assumption which we find overly restrictive and propose to circumvent the problem by introducing a modification to the LSW model (we call the new class of processes "LSW$_2$"). In Section 3.3, we derive Kolmogorov's formula for the one-step predicition error in the LSW$_2$ model. In Section 3.4, we investigate the behaviour of the (unsmoothed) local covariance estimator, used to estimate the entries of the approximate prediction matrix. We find that the estimator is asymptotically unbiased but inconsistent and thus needs to be smoothed. In Section 3.5, we propose an algorithm for choosing values of the nuisance parameters arising in the forecasting procedure (including the smoothing parameter for the covariance estimator). Finally, in Section 3.6, we demonstrate the performance of our forecasting algorithm on a time series of yearly values of the wind speed anomaly index in a specific region of the Pacific.

3.1 Forecasting by approximate MSPE minimisation

Assume that we have observed $X_{0,T}, X_{1,T}, \dots, X_{t-1,T}$ and want to predict $X_{t+h-1,T}$ for $h = 1, 2, \dots, T-t$. As we are only dealing with Gaussian LSW processes, it is legitimate to consider a *linear h-step predictor*

$$\hat{X}_{t+h-1,T} = \sum_{s=0}^{t-1} b^{(h)}_{t-1-s,T} X_{s,T}, \tag{3.1}$$

where, ideally, we would like the coefficients $\{b_{s,T}\}_{s=0}^{t-1}$ to minimise the MSPE:

$$\begin{aligned} &\mathbb{E}(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2 = \\ &\quad \mathbb{E}\left(\sum_{j=-1}^{-J(T)} \sum_{k\in\mathbb{Z}} \omega_{j,k;T} \left(\sum_{s=0}^{t-1} b^{(h)}_{t-1-s,T} \psi_{j,k-s} - \psi_{j,k-(t+h-1)}\right) \xi_{j,k}\right)^2 = \\ &\quad \sum_{j,k} \omega^2_{j,k;T} \left(\sum_{s=0}^{t-1} b^{(h)}_{t-1-s,T} \psi_{j,k-s} - \psi_{j,k-(t+h-1)}\right)^2 = \\ &\quad \sum_{m=0}^{T-1} \sum_{n=0}^{T-1} \tilde{b}^{(h)}_{m,T} \tilde{b}^{(h)}_{n,T} \sum_{j,k} \omega^2_{j,k;T} \psi_{j,k-m} \psi_{j,k-n}, \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} \tilde{b}^{(h)}_{n,T} &= b^{(h)}_{t-1-n,T} && \text{for} \quad n = 0, \dots, t-1 \\ \tilde{b}^{(h)}_{n,T} &= -1 && \text{for} \quad n = t-1+h \\ \tilde{b}^{(h)}_{n,T} &= 0 && \text{otherwise.} \end{aligned}$$

Defining

$$\begin{aligned}\tilde{\mathbf{b}}_T^{(h)} &= (\tilde{b}_{0,T}^{(h)}, \dots, \tilde{b}_{T-1,T}^{(h)}) \\ (\Sigma_T)_{m,n} &= \mathrm{Cov}(X_{m,T}, X_{n,T}) = \sum_{j,k} \omega_{j,k;T}^2 \psi_{j,k-m}\psi_{j,k-n},\end{aligned}$$

we can write (3.2) as a quadratic form

$$\mathbb{E}(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2 = \tilde{\mathbf{b}}_T^{(h)} \Sigma_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'. \tag{3.3}$$

However, (3.3) involves $\omega_{j,k;T}$'s which, as we said earlier, are non-identifiable (a given LSW process does not determine the sequence of $\omega_{j,k;T}$'s uniquely), and therefore cannot be estimated from the data. It is for this reason that in the LSW framework, it is both elegant and useful in practice to approximate quantities involving $\{\omega_{j,k;T}\}_{j,k}$ by ones involving $\{W_j(z)\}_j$ — an approach adopted in the original paper by Nason *et al.* (2000). We shall now investigate the possibility of approximating (3.3) by $\tilde{\mathbf{b}}_T^{(h)} \mathbf{B}_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'$, where

$$(\mathbf{B}_T)_{m,n} = \sum_{j=-1}^{-\infty} \sum_{k\in\mathbb{Z}} S_j\left(\frac{k}{T}\right) \psi_{j,k-m}\psi_{j,k-n}. \tag{3.4}$$

Note that both Σ_T and $\mathbf{B}_T$ are symmetric. We first show a result concerning the spectral norms of $\mathbf{B}_T$ and its inverse. Denote

$$\hat{\psi}_j(\omega) = \sum_n \psi_{j,n} \exp(i\omega n). \tag{3.5}$$

Lemma 3.1.1 *Let $\|\mathbf{A}\|$ denote the spectral norm of a quadratic matrix $\mathbf{A}$, and let $\overline{S}_j = \sup_z S_j(z)$ and $\underline{S}_j = \inf_z S_j(z)$. If*

$$\operatorname*{ess\,sup}_{\omega} \sum_j \overline{S}_j \left|\hat{\psi}_j(\omega)\right|^2 < \infty, \tag{3.6}$$

then $\|\mathbf{B}_T\|$ is bounded in T. Similarly, if

$$\operatorname*{ess\,inf}_{\omega} \sum_j \underline{S}_j \left|\hat{\psi}_j(\omega)\right|^2 > 0, \tag{3.7}$$

then $\left\|\mathbf{B}_T^{-1}\right\|$ is bounded in T.

Proof. As $\mathbf{B}_T$ is nonnegative definite, we have

$$\|\mathbf{B}_T\| \leq \left\|\mathbf{B}_T^{1/2}\right\|^2 = \sup_{\|x\|_2^2=1} x\mathbf{B}_T x', \tag{3.8}$$

and

$$\|\mathbf{B}_T^{-1}\| \leq \left\|\mathbf{B}_T^{-1/2}\right\|^2 = \sup_x \frac{x\mathbf{B}_T^{-1}x'}{xx'} = \sup_x \frac{xx'}{x\mathbf{B}_T x'} = \left(\inf_{\|x\|_2^2=1} x\mathbf{B}_T x'\right)^{-1}. \tag{3.9}$$

It remains to investigate the behaviour of the quadratic form $x\mathbf{B}_T x'$. Denote

$$\hat{x}(\omega) = \sum_n x_n \exp(i\omega n).$$

Simple algebra gives

$$\begin{aligned}
x\mathbf{B}_T x' &= \sum_{j,k} S_j(k/T) \left(\sum_n x_n \psi_{j,k-n}\right)^2 \\
&\leq \sum_j \overline{S}_j \sum_k \left(\sum_n x_n \psi_{j,k-n}\right)^2 \\
&= \frac{1}{2\pi} \sum_j \overline{S}_j \int_{-\pi}^{\pi} |\hat{x}(\omega)|^2 \left|\hat{\psi}_j(\omega)\right|^2 dw \\
&\leq \operatorname*{ess\,sup}_\omega \left\{\sum_j \overline{S}_j \left|\hat{\psi}_j(\omega)\right|^2\right\} \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{x}(\omega)|^2 d\omega \\
&= \operatorname*{ess\,sup}_\omega \left\{\sum_j \overline{S}_j \left|\hat{\psi}_j(\omega)\right|^2\right\} xx',
\end{aligned}$$

which proves the first part of the Lemma. Similar steps (with obvious modifications) are used to prove the second part. □

Proposition 3.1.1 *Let Σ_T and $\mathbf{B}_T$ arise from an LSW process satisfying*

$$T^{-1} \sum_{j=-J(T)}^{-1} C_j \mathcal{L}_j = o_T(1) \tag{3.10}$$

$$T \sum_{j=-J(T)-1}^{-\infty} \overline{S}_j = o_T(1) \tag{3.11}$$

and assumption (3.7). We have

$$\tilde{\mathbf{b}}_T^{(h)} \Sigma_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)' = \tilde{\mathbf{b}}_T^{(h)} \mathbf{B}_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)' (1 + o_T(1)). \tag{3.12}$$

Proof. We first consider an approximation by $\tilde{\mathbf{b}}_T^{(h)}\tilde{\mathbf{B}}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'$, where

$$(\tilde{\mathbf{B}}_T)_{m,n} = \sum_{j=-1}^{-J(T)} \sum_{k\in\mathbb{Z}} S_j\left(\frac{k}{T}\right)\psi_{j,k-m}\psi_{j,k-n}. \tag{3.13}$$

We have

$$\begin{aligned}&\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{B}}_T - \Sigma_T\right)\left(\tilde{\mathbf{b}}_T^{(h)}\right)' = \\ &\sum_{m,n}\sum_{j,k}\left(S_j\left(\frac{k}{T}\right) - \omega_{j,k;T}^2\right)\tilde{b}_{m,T}^{(h)}\tilde{b}_{n,T}^{(h)}\psi_{j,k-m}\psi_{j,k-n} \le \\ &\sum_{m,n}\sum_{j,k}\left|S_j\left(\frac{k}{T}\right) - \omega_{j,k;T}^2\right|\left|\tilde{b}_{m,T}^{(h)}\tilde{b}_{n,T}^{(h)}\psi_{j,k-m}\psi_{j,k-n}\right|.\end{aligned} \tag{3.14}$$

We know from Nason *et al.* (2000) that

$$\left|S_j\left(\frac{k}{T}\right) - \omega_{j,k;T}^2\right| \le \frac{C_j}{T}. \tag{3.15}$$

Thus, continuing from (3.14), we obtain

$$\begin{aligned}\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{B}}_T - \Sigma_T\right)\left(\tilde{\mathbf{b}}_T^{(h)}\right)' &\le T^{-1}\sum_{m,n}\sum_{j,k} C_j\left|\tilde{b}_{m,T}^{(h)}\tilde{b}_{n,T}^{(h)}\psi_{j,k-m}\psi_{j,k-n}\right|. \\ &= T^{-1}\sum_j C_j \sum_k\left(\sum_n\left|\tilde{b}_{n,T}^{(h)}\mathbb{I}_{\{\mathcal{L}_j>k-n\ge 0\}}\psi_{j,k-n}\right|\right)^2 \\ &\le T^{-1}\sum_j C_j\sum_n\left(\tilde{b}_{n,T}^{(h)}\right)^2\sum_k \mathbb{I}_{\{\mathcal{L}_j>k-n\ge 0\}}\sum_m \psi_{j,k-m}^2 \\ &= T^{-1}\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)'\sum_j C_j\mathcal{L}_j,\end{aligned}$$

using the Cauchy inequality and the property that $\sum_k \psi_{j,k}^2 = 1$. By assumption (3.10), we arrive at

$$\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{B}}_T - \Sigma_T\right)\left(\tilde{\mathbf{b}}_T^{(h)}\right)' = \tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)' o_T(1). \tag{3.16}$$

Let us now turn to the approximation of $\tilde{\mathbf{b}}_T^{(h)}\tilde{\mathbf{B}}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'$ by $\tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'$. Denoting

$$\hat{b}_T^{(h)}(\omega) = \sum_n \exp(i\omega n)\tilde{b}_{n,T}^{(h)}, \tag{3.17}$$

we have

$$\begin{aligned}
\tilde{\mathbf{b}}_T^{(h)}\left(\mathbf{B}_T-\tilde{\mathbf{B}}_T\right)\left(\tilde{\mathbf{b}}_T^{(h)}\right)' &= \sum_{m,n}\sum_{j=-J(T)-1}^{-\infty}\sum_k S_j(k/T)\psi_{j,k-m}\psi_{j,k-n}\tilde{b}_{m,T}^{(h)}\tilde{b}_{n,T}^{(h)} \\
&= \sum_{j=-J(T)-1}^{-\infty}\sum_k S_j(k/T)\left(\sum_n \tilde{b}_{n,T}^{(h)}\psi_{j,k-n}\right)^2 \\
&\leq \sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\sum_k\left(\sum_n \tilde{b}_{n,T}^{(h)}\psi_{j,k-n}\right)^2 \\
&= \frac{1}{2\pi}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\int_{-\pi}^{\pi}\left|\hat{b}_T^{(h)}(\omega)\right|^2\left|\hat{\psi}_j(\omega)\right|^2 d\omega \\
&\leq \sup_\omega\left|\hat{b}_T^{(h)}(\omega)\right|^2\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\hat{\psi}_j(\omega)\right|^2 d\omega \\
&\leq \left\|\tilde{\mathbf{b}}_T^{(h)}\right\|_{l_1}^2\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j \\
&\leq T\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)'\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j \\
&= \tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)' o_T(1), \qquad (3.18)
\end{aligned}$$

by assumption (3.11). Combining (3.16) and (3.18), we get

$$\tilde{\mathbf{b}}_T^{(h)}\left(\mathbf{B}_T-\Sigma_T\right)\left(\tilde{\mathbf{b}}_T^{(h)}\right)' = \tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)' o_T(1). \qquad (3.19)$$

Noting that

$$\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)' \leq \tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'\left\|\mathbf{B}_T^{-1}\right\| \qquad (3.20)$$

and using the second result of Lemma 3.1.1 completes the proof. □

Proposition 3.1.1 implies that there exists a sequence $d_T \downarrow 0$ such that

$$\tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'(1-d_T) \leq \tilde{\mathbf{b}}_T^{(h)}\Sigma_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)' \leq \tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'(1+d_T) \qquad (3.21)$$

which in turn means that

$$\begin{aligned}
\inf\tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'(1-d_T) &\leq \inf\tilde{\mathbf{b}}_T^{(h)}\Sigma_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)' \\
&= \mathrm{MSPE}(\hat{X}_{t+h-1,T}, X_{t+h-1,T}) \\
&\leq \inf\tilde{\mathbf{b}}_T^{(h)}\mathbf{B}_T\left(\tilde{\mathbf{b}}_T^{(h)}\right)'(1+d_T),
\end{aligned}$$

where the infimum has been taken w.r.t. $\tilde{b}_0^{(h)}, \ldots, \tilde{b}_{t-1}^{(h)}$. Thus, finding the h-step prediction error is asymptotically equivalent to minimising $\tilde{\mathbf{b}}_T^{(h)\prime} \mathbf{B}_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'$. Like in the classical (stationary) setting, the minimisation is performed by simple differentiation, yielding the system of prediction (or Yule-Walker) equations:

$$\sum_{n=0}^{t-1} \tilde{b}_n^{(h)} \sum_{j=-1}^{-\infty} \sum_{k \in \mathbb{Z}} S_j\left(\frac{k}{T}\right) \psi_{j,k-n} \psi_{j,k-m} = \sum_{j=-1}^{-\infty} \sum_{k \in \mathbb{Z}} S_j\left(\frac{k}{T}\right) \psi_{j,k-(t-1+h)} \psi_{j,k-m}, \tag{3.22}$$

for $m = 0, 1, \ldots, t-1$. Let $\mathbf{B}_{t,T}$ denote the matrix of this system. By a standard result in numerical analysis (Kress (1991), Theorem 5.3), the asymptotic stability of inversion of the system (3.22) is governed by the so-called condition number, defined by $\operatorname{cond}(\mathbf{B}_{t,T}) = \|\mathbf{B}_{t,T}\| \left\|\mathbf{B}_{t,T}^{-1}\right\|$: if $\operatorname{cond}(\mathbf{B}_{t,T}) \le C < \infty$ as $t \to \infty$, then the inversion is asymptotically numerically stable, i.e. "small" perturbations of the entries of $\mathbf{B}_{t,T}$ lead to "small" perturbations of the solution. Using identical reasoning as in Lemma 3.1.1, it can readily be shown that under assumptions (3.6) and (3.7) we have $\operatorname{cond}(\mathbf{B}_{t,T}) \le C < \infty$ as $T \to \infty$, uniformly in t.

Note that no assumption concerning the asymptotic behaviour of t has been made, and indeed, no such assumption is needed for the results of this section to hold.

It is interesting to observe that the entries of $(\mathbf{B}_{t,T})_{m,n}$ are not exactly asymptotic local covariances of $X_{t,T}$, as they cannot generally be represented in the form $c(z, m-n) = \sum_{j=-1}^{-\infty} S_j(z) \Psi_j(m-n)$ for any z. However, they can be approximated by e.g. $c((m+n)/2T, m-n)$ in the following sense:

$$\begin{aligned}
&\left|(\mathbf{B}_{t,T})_{m,n} - c\left(\frac{m+n}{2T}, m-n\right)\right| = \\
&\left|\sum_{j=-1}^{-\infty} \sum_{k \in \mathbf{Z}} \left(S_j\left(\frac{k}{T}\right) - S_j\left(\frac{m+n}{2T}\right)\right) \psi_{j,k-m} \psi_{j,k-n}\right| \le \\
&\sum_{j=-1}^{-\infty} \sum_{k \in \mathbf{Z}} \left|S_j\left(\frac{k}{T}\right) - S_j\left(\frac{m+n}{2T}\right)\right| \left|\psi_{j,k-m} \psi_{j,k-n}\right|.
\end{aligned} \tag{3.23}$$

We know from Nason *et al.* (2000) that $|S_j(z) - S_j(z+\delta)| \le L_j \delta / T$. Also, the ψ_j's are compactly supported, so $0 \le k - m < \mathcal{L}_j$ and $0 \le k - n < \mathcal{L}_j$, which implies

$0 \leq k - (m+n)/2 < \mathcal{L}_j$. Therefore, (3.23) can be bounded from above by

$$\begin{aligned} T^{-1} \sum_{j=-1}^{-\infty} L_j \mathcal{L}_j \sum_{k \in \mathbb{Z}} |\psi_{j,k-m} \psi_{j,k-n}| &\leq T^{-1} \sum_{j=-1}^{-\infty} L_j \mathcal{L}_j \sum_k \psi_{j,k}^2 \\ &= T^{-1} \sum_{j=-1}^{-\infty} L_j \mathcal{L}_j \\ &= O(T^{-1}), \end{aligned}$$

using the Cauchy inequality in the first step, the property that $\sum_k \psi_{j,k}^2 = 1$ in the second step, and the definition of the LSW process in the final one (note that $\mathcal{L}_j = O(2^{-j})$). Thus, the entries of $\mathbf{B}_{t,T}$ are uniformly close to the corresponding asymptotic local covariances.

If the second-order structure of the process was known, the system of prediction equations (3.22) could be solved e.g. using the innovations algorithm (see Brockwell & Davis (1987), Section 5.2) to yield the prediction coefficients $\{\tilde{b}_n^{(h)}\}_{n=0}^{t-1}$. However, in practice the second-order structure needs to be estimated from the data: see Section 3.4 for details of the estimation procedure.

3.2 A closer look at the results of Section 3.1

In this section, we investigate whether the assumptions of Lemma 3.1.1 and Proposition 3.1.1 can be regarded as "restrictive" and if so, what can be done to relax them.

3.2.1 Assumptions of Lemma 3.1.1

First of all, note that assumptions (3.6) and (3.7) are "LSW counterparts" of the classical assumptions from stationary time series theory that the spectral density be bounded from above and bounded away from zero (respectively). Indeed, let X_t be a *stationary* LSW process with wavelet spectrum $\{S_j\}_j$, covariance $\gamma(\tau)$ and spectral density $f(\omega)$. We have

$$\int_{-\pi}^{\pi} f(\omega)\exp(i\omega n)d\omega =$$
$$\int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{\tau}\gamma(\tau)\exp(-i\omega\tau)\exp(i\omega n)d\omega =$$
$$\int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{\tau}\sum_{j} S_j\Psi_j(\tau)\exp(-i\omega\tau)\exp(i\omega n)d\omega =$$
$$\int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{\tau}\sum_{j} S_j\sum_{k}\psi_{j,k}\psi_{j,k+\tau}\exp(i\omega k)\exp(-i\omega(\tau+k))\exp(i\omega n)d\omega =$$
$$\int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{j} S_j\left|\hat{\psi}_j(\omega)\right|^2\exp(i\omega n)d\omega$$

for all n, so that

$$f(\omega) = \frac{1}{2\pi}\sum_{j} S_j\left|\hat{\psi}_j(\omega)\right|^2 \quad \text{a.e.} \tag{3.24}$$

and assumptions (3.6) and (3.7) simplify as

$$0 < \operatorname*{ess\,inf}_{\omega} f(\omega) \le \operatorname*{ess\,sup}_{\omega} f(\omega) < \infty.$$

Note that in the LSW case, it is necessary to use "ess inf" instead of "inf", due to the following fact:

$$\begin{aligned}\inf_{\omega}\sum_{j=-1}^{-\infty} S_j(z)\left|\hat{\psi}_j(\omega)\right|^2 &= \sum_{j=-1}^{-\infty} S_j(z)\left|\hat{\psi}_j(0)\right|^2 \\ &= \sum_{j=-1}^{-\infty} S_j(z)\left|\sum_{k}\psi_{j,k}\right|^2 \\ &= 0,\end{aligned} \tag{3.25}$$

using the property that $\sum_k \psi_{j,k} = 0$.

There arises a natural question whether there exist LSW processes for which (3.7) is satisfied, even though, as we have shown in (3.25), the same condition with "ess inf" replaced by "inf" is not satisfied by *any* LSW process. The (reassuring) answer is yes: S. Van Bellegem shows in Fryzlewicz *et al.* (2003) that standard white noise is an LSW process with $S_j = 2^j$, so, by (3.24), we must have

$$1 = \sum_{j} 2_j\left|\hat{\psi}_j(\omega)\right|^2 \quad \text{a.e.} \tag{3.26}$$

for any system of compactly supported Daubechies' wavelets ψ_j, and, clearly, (3.7) is then satisfied. Let us now give an example of a class of LSW processes for which (3.7) does not hold. We first define sparse LSW processes.

Definition 3.2.1 *An LSW process $X_{t,T}$ with spectrum $\{S_j(z)\}_j$ is said to be* sparse *if $S_j(z) \equiv 0$ for all j except for a finite set.*

Proposition 3.2.1 *No sparse LSW process satisfies (3.7).*

Proof. Let $D = \{j : S_j(z) \not\equiv 0\}$. Being a Fourier transform of a finite-length vector, $\hat{\psi}_j(\omega)$ is continuous for all j, so $\sum_{j\in D} S_j(z) \left|\psi_j(\omega)\right|^2$ is continuous as a finite sum of continuous functions. Therefore,

$$\operatorname{ess\,inf}_{\omega} \sum_{j\in D} S_j(z) \left|\psi_j(\omega)\right|^2 = \inf_{\omega} \sum_{j\in D} S_j(z) \left|\psi_j(\omega)\right|^2 = 0, \tag{3.27}$$

and (3.7) is violated. □

This is certainly bad news from the point of view of the philosophy of LSW modelling. Indeed, sparse LSW processes which have an economical representation in the model and are therefore appealing, are "badly behaved" as far as forecasting is concerned: for them, the system of prediction equations (3.22) cannot be solved numerically in a stable manner. One of the avenues for future research might be to investigate how this situation can be remedied by modifying the definition of an LSW process.

3.2.2 Assumptions of Proposition 3.1.1

A purely technical assumption (3.10) controls the evolution of the sequence $\{C_j\}_j$. An assumption like this is inevitable in the context of approximating the finite-sample MSPE $\tilde{\mathbf{b}}_T^{(h)} \Sigma_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'$ by $\tilde{\mathbf{b}}_T^{(h)} \tilde{\mathbf{B}}_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'$.

On the other hand, assumption (3.11), controlling the "tail behaviour" of the sequence $\{\overline{S}_j\}_j$, is extremely restrictive. Indeed, even the white noise process does not satisfy it: noting that $J(T) = O(\log_2(T))$, we obtain

$$T \sum_{j=-J(T)-1}^{-\infty} 2^j = TO(T^{-1}) = O(1) \neq o_T(1). \tag{3.28}$$

By inspecting the proof of Proposition 3.1.1, it is easy to see why there is a need to control the tail behaviour of $\{\overline{S}_j\}_j$. The underlying reason is that in the definition of an LSW process, $X_{t,T}$ is only built of wavelets at the $J(T)$ finest scales (i.e. the summation over j only goes from -1 to $-J(T)$), whereas the asymptotic quantities such as $c(z, \tau)$, or indeed $\tilde{\mathbf{b}}_T^{(h)} \mathbf{B}_T \left(\tilde{\mathbf{b}}_T^{(h)}\right)'$, typically involve the wavelet spectrum at *all* scales, i.e. $\{S_j(z)\}_{j=-1}^{-\infty}$. Therefore, without controlling the "tail" of the sequence $\{S_j(z)\}_{j=-1}^{-\infty}$ in one way or another, we cannot hope to achieve the desired rates of convergence.

However, no assumption controlling the tail of $\{S_j(z)\}_{j=-1}^{-\infty}$ is made in the original paper by Nason *et al.* (2000). To illustrate the implications of this fact, note that the result of Proposition 2.11 from Nason *et al.* (2000) does not formally hold without such an assumption. Let us first recall the statement of the proposition.

Proposition 3.2.2 (Nason *et al.* (2000)) *As $T \to \infty$, uniformly in $\tau \in \mathbb{Z}$ and $z \in (0, 1)$, $|c_T(z, \tau) - c(z, \tau)| = O(T^{-1})$.*

It is easy to find a counterexample to the above proposition. Consider a stationary process with $\omega_{j,k;T} = W_j(z) = -1/j$. This process is LSW in the sense of Nason *et al.* (2000), and

$$|c_T(z, 0) - c(z, 0)| = \sum_{-J(T)-1}^{-\infty} \frac{1}{j^2}, \tag{3.29}$$

which behaves like

$$\int_{\log(T)}^{\infty} \frac{1}{x^2} dx = -\frac{1}{x}\Big|_{\log(T)}^{\infty} = \log^{-1}(T) \neq O(T^{-1}), \tag{3.30}$$

and this disproves the authors' claim.

An easy way to avoid such "tail considerations" altogether is to assume that the summation over j in the definition of an LSW process ranges from -1 to $-\infty$, even in the finite sample situation. Indeed, this is also implicitly done in the proof of Proposition 2.11 in Nason *et al.* (2000) (thereby enabling the authors to achieve the claimed rate of $O(T^{-1})$).

For completeness, we give the amended definition below. Note the subscript in "LSW$_2$".

Definition 3.2.2 *A triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$, for $T = 1, 2, \ldots$, is in the class of* LSW_2 *processes if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \omega_{j,k;T} \psi_{j,k}(t) \xi_{j,k}, \tag{3.31}$$

where $\psi_{j,k}(t)$ are nondecimated discrete wavelet vectors, $\omega_{j,k;T}$ are real constants, and $\{\xi_{j,k}\}_{j,k}$ are zero-mean orthonormal identically distributed random variables. Also, we assume that for each $j \leq -1$, there exists a Lipschitz function $W_j(z) : [0,1] \to \mathbb{R}$ such that

- $\sum_{j=-\infty}^{-1} |W_j|^2 < \infty$,
- *the Lipschitz constants L_j satisfy*

$$\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty, \tag{3.32}$$

- *there exists a sequence of constants C_j satisfying $\sum_{j=-\infty}^{-1} C_j < \infty$ such that, for each T,*

$$\sup_{k=0,\ldots,T-1} |\omega_{j,k;T} - W_j(k/T)| \leq C_j/T \quad \forall j. \tag{3.33}$$

In other words, *all* "building blocks" ψ_j are included in the construction of $X_{t,T}$, even in the finite sample case. What lends credibility to the above definition is the fact that similar approach was adopted in Dahlhaus' theory of locally stationary processes, where the entire set of building blocks $\{\exp(i\omega t)\}_{\omega \in (-\pi,\pi]}$ was used to construct $X_{t,T}$, even in the finite sample situation.

Note that if the word "LSW" is replaced by "LSW$_2$" in Proposition 3.1.1, then assumption (3.11) becomes unnecessary. Lemma 3.1.1 holds for LSW$_2$ processes in an unchanged form.

On a final note, we come back to the proof of Proposition 3.1.1. In the derivation of (3.18), is tempting to write

$$
\begin{aligned}
&\frac{1}{2\pi}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\int_{-\pi}^{\pi}\left|\hat{b}_T^{(h)}(\omega)\right|^2\left|\hat{\psi}_j(\omega)\right|^2 d\omega = \\
&\quad\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\hat{b}_T^{(h)}(\omega)\right|^2\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2 d\omega \leq \\
&\quad\operatorname*{ess\,sup}_{\omega}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\hat{b}_T^{(h)}(\alpha)\right|^2 d\alpha = \\
&\quad\operatorname*{ess\,sup}_{\omega}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2\tilde{\mathbf{b}}_T^{(h)}\left(\tilde{\mathbf{b}}_T^{(h)}\right)',
\end{aligned}
\tag{3.34}
$$

in the hope that

$$
\operatorname*{ess\,sup}_{\omega}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2 = o_T(1),
$$

as certainly would be the case if $\sum_{j=-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2$ was continuous (by Dini's theorem). However, we showed in Section 3.2.1 that this need not be the case. Indeed, we have

$$
\begin{aligned}
&\operatorname*{ess\,sup}_{\omega}\sum_{j=-J(T)-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2 \geq \\
&\quad\operatorname*{ess\,inf}_{\omega}\sum_{j=-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2 - \inf_{\omega}\sum_{j=-1}^{-J(T)}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2 = \\
&\quad\operatorname*{ess\,inf}_{\omega}\sum_{j=-1}^{-\infty}\overline{S}_j\left|\hat{\psi}_j(\omega)\right|^2,
\end{aligned}
$$

and the last quantity can be strictly positive, e.g. for white noise. This shows that transformation step (3.34) would not be helpful in proving Proposition 3.1.1.

3.3 Kolmogorov's formula for LSW_2 processes

In this section, we state and prove Kolmogorov's formula for the one-step MSPE in the LSW_2 framework. The only difference with the LSW setting is that the sums over j in the finite-sample quantities go from -1 to $-\infty$ and not to $-J(T)$.

We first recall the statement of Kolmogorov's formula in the classical stationary setting (Brockwell & Davis (1987), Theorem 5.8.1).

Theorem 3.3.1 *Let $\{X_t\}$ be a real-valued zero-mean stationary process with spectral density function f. The one-step MSPE of $\{X_t\}$ is*

$$\sigma^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\log(2\pi f(\omega))d\omega\right\}. \tag{3.35}$$

An analogous formula was derived by Dahlhaus in the locally stationary model (Dahlhaus (1996b), Theorem 3.2 (i)). We follow his method of proof here; however, some important modifications are needed due to the fact that the building blocks in the LSW model are wavelets and not Fourier exponentials. We first introduce some essential notation. The observation domain $\{0,\ldots,T-1\}$ is divided into overlapping blocks I_m of length N with shift S (assume that both T and N are multiples of S). At the edges the length of the blocks is reduced, but the shift kept so that each observation is contained in exactly N/S blocks:

$$I_m = \begin{cases} [0,\ldots,mS-1] & m=1,\ldots,N/S \\ [mS-N,\ldots,mS-1] & m=N/S+1,\ldots,T/S \\ [mS-N,\ldots,T-1] & m=T/S+1,\ldots,(T+N)/S-1. \end{cases} \tag{3.36}$$

If T is not divisible by S then we "clip" the final blocks in the natural way (note that in this case we still have each observation contained in exactly N/S blocks). Let $M=(T+N)/S-1$ be the total number of blocks and let t_m be an arbitrary point $\in I_m$. For each $m=1,\ldots,M$, we define a $T\times T$ matrix

$$\left(\mathbf{D}_T^{(m)}\right)_{u,s} = \sum_{j=-1}^{-\infty} S_j\left(\frac{t_m}{T}\right)\Psi_j(u-s)\mathbb{I}_{\{u,s\in I_m\}}, \tag{3.37}$$

where the indices u,s go from 0 to $T-1$. Define further

$$\mathbf{D}_T = \frac{S}{N}\sum_{m=1}^{M}\mathbf{D}_T^{(m)} \tag{3.38}$$

and

$$c^*(k) = \sup_z |c(z,k)|. \tag{3.39}$$

Consider the following set of assumptions.

$$\begin{aligned} N &\to \infty && (3.40)\\ S/N &\to 0 && (3.41)\\ N^2/T &\to 0 && (3.42)\\ \sum_{k=1}^{\infty} c^*(k) &< \infty && (3.43)\\ \sum_{j=-1}^{-\infty} \mathcal{L}_j(C_j + L_j\mathcal{L}_j) &< \infty && (3.44)\end{aligned}$$

(3.40) – (3.42) are purely technical assumptions concerning the behaviour of S and N in relation to T. Assumption (3.43) is a non-stationary equivalent of the short memory property. Finally, assumption (3.44) controls the speed of convergence of the sequences $\{C_j\}$ and $\{L_j\}$ and is similar to (3.10).

We begin with the following lemma.

Lemma 3.3.1 *Let x be a row vector of length T. Under assumptions (3.40) – (3.44), we have*

$$x\left(\Sigma_T - \mathbf{D}_T\right)x' = xx'o_T(1). \tag{3.45}$$

Proof. Define

$$\left(\Sigma_T^{(m)}\right)_{u,s} = \left(\Sigma_T\right)_{u,s\in I_m}. \tag{3.46}$$

We have

$$\begin{aligned} x\left(\Sigma_T - \mathbf{D}_T\right)x' &= x\left(\frac{S}{N}\sum_{m=1}^{M}\Sigma_T^{(m)} - \mathbf{D}_T^{(m)}\right)x' \\ &+ \sum_{k,l=0}^{T/S-1}\min\left\{|k-l|\frac{S}{N},1\right\}\sum_{u,s=0}^{S-1} x_{kS+u}\left(\Sigma_T\right)_{kS+u,lS+s}x_{lS+s} \end{aligned} \tag{3.47}$$

We will first show that the second term tends to zero. Replace $\left(\Sigma_T\right)_{u,s}$ by $c((u+$

$s)/2T, u-s)$. The second term is bounded by

$$\sum_{d=1}^{T/S-1} \min\left\{d\frac{S}{N}, 1\right\} \sum_{\substack{u,s=0 \\ (d-1)S<|u-s|\leq dS}}^{T-1} \left| x_u c\left(\frac{u+s}{2T}, u-s\right) x_s \right| + R \leq$$

$$2xx' \sum_{d=1}^{T/S-1} \min\left\{d\frac{S}{N}, 1\right\} \sum_{k=(d-1)S+1}^{dS} c^*(k) + R \leq$$

$$2xx' \left(\frac{S+\sqrt{N}}{N} \sum_{k=1}^{\infty} c^*(k) + \sum_{k>\sqrt{N}} c^*(k) \right) + R,$$

and the first term in the above sum is of order $xx'o_T(1)$ by assumptions (3.40), (3.41) and (3.43). Let us now turn to the remainder R. We have

$$R \leq \sum_{u,s=0}^{T-1} \left| x_u x_s \sum_{j,k} \left(\omega^2_{j,k;T} - S_j\left(\frac{u+s}{2T}\right)\right) \psi_{j,k-s}\psi_{j,k-u} \right|, \tag{3.48}$$

and, using exactly the same technique as in Proposition 3.1.1, it can be shown that $R = xx'O(T^{-1})$ under assumption (3.44).

We now consider the main term in (3.47). Denote by $\underline{I}_m$ and $\overline{I}_m$, respectively, the initial and final indices in the segment I_m. We have

$$x\left(\frac{S}{N}\sum_{m=1}^{M}\Sigma_T^{(m)} - \mathbf{D}_T\right)x' =$$

$$\frac{S}{N}\sum_{m=1}^{M}\sum_{j,k}\left(\omega^2_{j,k;T} - S_j\left(\frac{t_m}{T}\right)\right)\left(\sum_u \psi_{j,k-u}x_u \mathbb{I}_{\{u\in I_m\}}\right)^2 \leq$$

$$\frac{S}{N}\sum_{m=1}^{M}\sum_{j}\sum_{k=\underline{I}_m}^{\overline{I}_m+\mathcal{L}_j-1} \frac{C_j + L_j(\mathcal{L}_j+N)}{T}\left(\sum_u x_u^2 \mathbb{I}_{\{u\in I_m\}}\right) \leq$$

$$\frac{S}{N}\sum_{m=1}^{M}\left(\sum_u x_u^2 \mathbb{I}_{\{u\in I_m\}}\right)\sum_j \frac{(C_j + L_j(\mathcal{L}_j+N))(\mathcal{L}_j+N)}{T} =$$

$$xx'\sum_j \frac{(C_j + L_j(\mathcal{L}_j+N))(\mathcal{L}_j+N)}{T}, \tag{3.49}$$

where the last equality holds because by construction each x_u is contained in exactly N/S segments I_m. By assumptions (3.44) and (3.42), the above is of order $xx'o_T(1)$, which completes the proof. □

To derive Kolmogorov's formula, we also need another (similar) lemma.

Lemma 3.3.2 *Suppose that assumption (3.44) holds and that there exists a t^* such that $x_u = 0$ for all $u \notin \{t^*, \ldots, t^* + L\}$. Then for each $\tilde{t} \in \{t^*, \ldots, t^* + L\}$*

$$x\Sigma_T x' = \sum_j S_j\left(\frac{\tilde{t}}{T}\right)\sum_k\left(\sum_{u=t^*}^{t^*+L} x_u\psi_{j,k-u}\right)^2 + xx'O\left(\frac{L^2}{T}\right). \tag{3.50}$$

The proof is completely analogous to the part of the proof of Lemma 3.3.1 leading to the bound for the main term, i.e. formula (3.49).

Before moving on to the statement of Kolmogorov's formula for LSW_2 processes, we present an interesting technique for bounding the spectral norms of Σ_T and its inverse. Suppose that the assumptions of Lemma 3.3.1 hold and that x is a row vector of length T. As Σ_T is nonnegative definite, its spectral norm is bounded from above by

$$\begin{aligned}
\sup_{\|x\|_2=1} x\Sigma_T x' &= \sup_{\|x\|_2=1} x\frac{S}{N}\sum_{m=1}^{M}\mathbf{D}_T^{(m)}x' + o_T(1)\\
&= \sup_{\|x\|_2=1}\frac{S}{N}\sum_{m=1}^{M}\sum_{j<0}S_j\left(\frac{t_m}{T}\right)\sum_k\left(\sum_u x_u\psi_{j,k-u}\mathbb{I}_{\{u\in I_m\}}\right)^2 + o_T(1)\\
&= \sup_{\|x\|_2=1}\frac{S}{2\pi N}\sum_{m=1}^{M}\int_{-\pi}^{\pi}\sum_{j<0}S_j\left(\frac{t_m}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\times\\
&\times\left|\sum_u x_u\mathbb{I}_{\{u\in I_m\}}e^{-i\omega u}\right|^2 d\omega + o_T(1)\\
&\le \operatorname*{ess\,sup}_{z,\omega}\sum_{j<0}S_j(z)\left|\hat{\psi}_j(\omega)\right|^2\sup_{\|x\|_2=1}xx' + o_T(1)\\
&= \operatorname*{ess\,sup}_{z,\omega}\sum_{j<0}S_j(z)\left|\hat{\psi}_j(\omega)\right|^2 + o_T(1).
\end{aligned}$$

In the same way, it can be shown that

$$\inf_{\|x\|_2=1} x\Sigma_T x' \ge \operatorname*{ess\,inf}_{z,\omega}\sum_{j<0}S_j(z)\left|\hat{\psi}_j(\omega)\right|^2 + o_T(1). \tag{3.51}$$

As we can always choose S and N with the properties required in (3.40) – (3.42), the only restrictive assuptions needed for the above derivation to be valid are (3.43) and (3.44). With these assumptions, $\|\Sigma_T\|$ is bounded in T if

$$\operatorname*{ess\,sup}_{z,\omega}\sum_{j<0}S_j(z)\left|\hat{\psi}_j(\omega)\right|^2 < \infty, \tag{3.52}$$

and $\|\Sigma_T^{-1}\|$ is bounded in T if

$$\operatorname*{ess\,inf}_{z,\omega} \sum_{j<0} S_j(z) \left|\hat{\psi}_j(\omega)\right|^2 > 0. \tag{3.53}$$

Two remarks are in order.

1. Observe that the short-memory property (3.43) implies (3.52). Indeed,

$$\begin{aligned}
\sum_j S_j(z) \left|\hat{\psi}_j(\omega)\right|^2 &= \sum_j S_j(z) \sum_k \psi_{j,k} \exp(i\omega k) \sum_n \psi_{j,n} \exp(-i\omega n) \\
&= \sum_j S_j(z) \sum_k \psi_{j,k} \sum_\tau \psi_{j,k+\tau} \exp(-i\omega\tau) \\
&= \sum_\tau \exp(-i\omega\tau) \sum_j S_j(z) \Psi_j(\tau) \\
&\leq \sum_\tau c^*(\tau).
\end{aligned}$$

   In the classical stationary setting, the analogous well known fact says that the absolute summability of the covariance implies the boundedness of the spectral density from above.

2. Note that we could also bound the norms of Σ_T and its inverse using Lemma 3.1.1 and Proposition 3.1.1: first by approximating $x\Sigma_T x'$ by $x\mathbf{B}_T x'$, and then using the boundedness of the norms of $\mathbf{B}_T$ and its inverse (in other words, we could use $\mathbf{B}_T$ instead of $\mathbf{D}_T$). For the approximation by $x\mathbf{B}_T x'$ to be valid, we would need an LSW$_2$ version of assumption (3.10):

$$\sum_{j=-\infty}^{-1} C_j \mathcal{L}_j < \infty \tag{3.54}$$

   (recall that (3.11) is not required in the LSW$_2$ setting). With assumption (3.54), $\|\Sigma_T\|$ is bounded in T if (3.6) holds, and $\|\Sigma_T^{-1}\|$ is bounded in T if (3.7) holds.

We are now in a position to state and prove Kolmogorov's formula for LSW$_2$ processes.

Theorem 3.3.2 *Let $X_{t,T}$ be an* LSW_2 *process satisfying assumptions (3.43), (3.44) and (3.53). Let $\hat{X}_{t,T}$ be the best linear predictor of $X_{t,T}$ given $X_{0,T}, \ldots, X_{t-1,T}$. Then*

$$\mathbb{E}\left(\hat{X}_{t,T} - X_{t,T}\right)^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + o_T(1). \tag{3.55}$$

Proof. Let x be a row vector of length T such that $x_0, \ldots, x_{t-1}$ are arbitrary, $x_t = -1$, and $x_{t+1}, \ldots, x_{T-1} = 0$. By Lemma 3.3.1, we have

$$\begin{aligned} x\Sigma_T x' &= x\mathbf{D}_T x' + xx' o_T(1) \\ &= \frac{S}{N}\sum_{m=1}^{M}\sum_{j<0} S_j\left(\frac{t_m}{T}\right)\sum_k\left(\sum_u x_u \psi_{j,k-u}\mathbb{I}_{\{u\in I_m\}}\right)^2 + xx' o_T(1). \end{aligned}$$

Let $\mathcal{M}_t = \{m : t \in I_m\}$. For $m \in \mathcal{M}_t$, we set $t_m = t$. The above expression is bounded from below by

$$\frac{S}{N}\sum_{m\in\mathcal{M}_t}\sum_{j<0} S_j\left(\frac{t_m}{T}\right)\sum_k\left(\sum_u x_u \psi_{j,k-u}\mathbb{I}_{\{u\in I_m\}}\right)^2 + xx' o_T(1). \tag{3.56}$$

Each of the sums over j represents the one-step prediction error for a stationary time series with spectral density

$$f(\omega) = \frac{1}{2\pi}\sum_j S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2. \tag{3.57}$$

There are exactly S/N such sums. By classical Kolmogorov's formula, each of them is bounded from below by

$$\exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\}.$$

Therefore, the lower bound is

$$x\Sigma_T x' \geq \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + xx' o_T(1). \tag{3.58}$$

We now turn to the remainder $xx' o_T(1)$. We have

$$\left(\frac{x}{\|x\|_2}\right)\left(\frac{x}{\|x\|_2}\right)' \leq \left(\frac{x}{\|x\|_2}\right)\Sigma_T\left(\frac{x}{\|x\|_2}\right)'\left\|\Sigma_T^{-1}\right\| \leq \|\Sigma_T\|\left\|\Sigma_T^{-1}\right\| \tag{3.59}$$

and thus

$$xx' \leq x\Sigma_T x' \left\|\Sigma_T^{-1}\right\| \leq xx' \left\|\Sigma_T\right\| \left\|\Sigma_T^{-1}\right\|. \tag{3.60}$$

$\|\Sigma_T\|$ is bounded by assumption (3.43) (see (3.52) and Remark 1. above), and $\left\|\Sigma_T^{-1}\right\|$ is bounded by assumption (3.53). Therefore, $xx' = O(x\Sigma_T x')$ and, continuing from (3.58),

$$x\Sigma_T x'(1 - o_T(1)) \geq \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\}, \tag{3.61}$$

which finally yields

$$\begin{aligned}
\mathbb{E}\left(\hat{X}_{t,T} - X_{t,T}\right)^2 &= \inf_x x\Sigma_T x' \\
&\geq \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + o_T(1).
\end{aligned}$$

To obtain an upper bound, we set $\bar{t} = \min(t, L)$ with $L^2/T \to 0$. Let $y_0^*, \dots, y_{\bar{t}-1}^*$ be the coefficients for the best linear predictor for a stationary process ($\bar{t}$ observations) with spectral density (3.57). Set $x_j = y_{j+\bar{t}-t}^*$ for $j = t - \bar{t}, \dots, t-1$, and $x_t = -1$, all the other components of x being zero. By Lemma 3.3.2, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{X}_{t,T} - X_{t,T}\right)^2 &\leq x\Sigma_T x' \\
&= \sum_j S_j\left(\frac{t}{T}\right)\sum_k\left(\sum_{u=t-\bar{t}}^{t} x_u \psi_{j,k-u}\right)^2 + xx' o_T(1).
\end{aligned}$$

Since the sum over j is the one-step prediction error for the stationary process with spectral density (3.57), we have

$$x\Sigma_T x' = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + xx' o_T(1). \tag{3.62}$$

By identical reasoning as in the derivation of the lower bound, this reduces to

$$x\Sigma_T x' = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + o_T(1) \tag{3.63}$$

and

$$\mathbb{E}\left(\hat{X}_{t,T} - X_{t,T}\right)^2 \leq \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j<0} S_j\left(\frac{t}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\right) d\omega\right\} + o_T(1), \tag{3.64}$$

which completes the proof. $\square$

3.4 Estimation of the approximating matrix $\mathbf{B}_T$

In order to perform forecasting in practice, we need to be able to estimate the entries of the approximating matrix $\mathbf{B}_T$. As was mentioned in Section 2.2.4, Nason *et al.* (2000) used the wavelet periodogram, defined in (2.36), to perform inference on the wavelet spectrum $\{S_j(z)\}$ in the LSW model. We also base our estimators of $\mathbf{B}_T$ in the LSW$_2$ framework on the wavelet periodogram.

Note that the wavelet periodogram $I_{t,T}^{(j)}$ is a function of $X_{t+1-\mathcal{L}_j}, \ldots, X_{t,T}$. If $t < \mathcal{L}_j - 1$, then one possibility is to assume periodicity in the data; however, for the theoretical results of this section to hold, we set $I_{t,T}^{(j)} = I_{\mathcal{L}_j-1,T}^{(j)}$ for $t = 0, \ldots, \mathcal{L}_j - 2$. We only compute the periodogram down to scale $j = -J(t)$. Throughout this section, we assume that t/T remains constant as $T \to \infty$.

As was shown in the final part of Section 3.1, the entries of $\mathbf{B}_T$ tend to the corresponding local autocovariances at the uniform rate of $O(T^{-1})$. Therefore, asymptotically, estimating the entries of $\mathbf{B}_T$ is equivalent to estimating the local autocovariance structure of the process. In constructing the estimator of $c(z, \tau)$, we first consider the case $\tau = 0$ (local variance). We define our estimator $\hat{c}(k/T, 0)$ as

$$\hat{c}\left(\frac{k}{T}, 0\right) = \sum_{j=-J(t)}^{-1} 2^j I_{k,T}^{(j)}, \quad \text{for} \quad k = 0, \ldots, t-1. \tag{3.65}$$

The extension of (3.65) to $\tau \neq 0$ uses the infinite matrix A, defined by (2.34). The invertibility of A for Haar wavelets was proved in Nason *et al.* (2000). Even though numerical results suggest that A is also invertible for other Daubechies' wavelets, no proof of this conjecture has as yet been established. For $\tau \neq 0$, we define $\hat{c}(k/T, \tau)$ as follows:

$$\hat{c}\left(\frac{k}{T}, \tau\right) = \sum_{j=-J(t)}^{-1} \left(\sum_{l=-\infty}^{-1} (A^{-1})_{j,l} \Psi_l(\tau) \right) I_{k,T}^{(j)}. \tag{3.66}$$

Before we analyse some properties of $\hat{c}(k/T, \tau)$, we quote the following lemma from Fryzlewicz *et al.* (2003).

Lemma 3.4.1 *The matrix A defined in (2.34) has the following properties:*

$$\sum_{j=-1}^{-\infty} 2^j A_{i,j} = 1 \quad (3.67)$$

$$\sum_{j=-1}^{-\infty} (A^{-1})_{i,j} = 2^i \quad (3.68)$$

$$\sum_{j=-1}^{-\infty} \left|(A^{-1})_{i,j}\right| = O(2^{i/2}), \quad (3.69)$$

where (3.68) and (3.69) only apply to Haar wavelets.

The proofs of (3.67) and (3.68) rely on the following result, due to S. Van Bellegem (Fryzlewicz *et al.* (2003), Lemma B.2):

$$\sum_j 2^j \Psi_j(\tau) = \delta_\tau, \quad (3.70)$$

where δ_τ is the Kronecker delta. The proof of (3.68) is also due to S. Van Bellegem.

The following proposition concerns the asymptotic behaviour of the first two moments of $\hat{c}(k/T, 0)$:

Proposition 3.4.1 *If (3.43) holds, then the estimator (3.65) satisfies*

$$\mathbb{E}\left\{\hat{c}\left(\frac{k}{T}, 0\right)\right\} = c\left(\frac{k}{T}, 0\right) + O(T^{-1}\log(T)). \quad (3.71)$$

If, in addition, the increment process $\{\xi_{j,k}\}$ is Gaussian, then

$$\mathrm{Var}\left\{\hat{c}\left(\frac{k}{T}, 0\right)\right\} = 2\sum_{i,j=-J(t)}^{-1} 2^{i+j}\left(\sum_\tau c(k/T,\tau)\sum_n \psi_{i,n-\tau}\psi_{j,n}\right)^2 + O(T^{-1}). \quad (3.72)$$

Proof. We will first show

$$\mathrm{cov}\left(\sum_s X_{s,T}\psi_{i,k-s}, \sum_s X_{s,T}\psi_{j,k-s}\right) = \sum_\tau c(k/T,\tau)\sum_n \psi_{i,n-\tau}\psi_{j,n} + O\left(2^{-(i+j)/2}T^{-1}\right). \quad (3.73)$$

We have

$$\mathrm{cov}\left(\sum_s X_{s,T}\psi_{i,k-s}, \sum_s X_{s,T}\psi_{j,k-s}\right) = \sum_{l,u}\left(S_l\left(\frac{k}{T}\right) + O\left(\frac{C_l + L_l(u-k)}{T}\right)\right)\sum_{s,t}\psi_{l,s-u}\psi_{j,k-s}\psi_{l,t-u}\psi_{i,k-t}.$$

Using $\mathcal{L}_j = O(M2^{-j})$ in the first step, and the Cauchy inequality in the second one, we bound the remainder as follows:

$$\left|\sum_{l,u} O\left(\frac{C_l + L_l(u-k)}{T}\right) \sum_{s,t} \psi_{l,s-u}\psi_{j,k-s}\psi_{l,t-u}\psi_{i,k-t}\right| \leq$$
$$\sum_l \frac{C_l + ML_l(2^{-l} + \min(2^{-i}, 2^{-j}))}{T} \sum_u \left|\sum_{s,t} \psi_{l,s-u}\psi_{j,k-s}\psi_{l,t-u}\psi_{i,k-t}\right| \leq$$
$$\sum_l \frac{C_l + ML_l(2^{-l} + 2^{-i/2}2^{-j/2})}{T} (A_{l,j})^{1/2}(A_{l,i})^{1/2} =$$
$$\frac{2^{-(i+j)/2}}{T}\left\{\sum_l (C_l + ML_l 2^{-l})2^{(i+j)/2}(A_{l,j})^{1/2}(A_{l,i})^{1/2} + \right.$$
$$\left. + \sum_l ML_l(A_{l,j})^{1/2}(A_{l,i})^{1/2}\right\} =$$
$$\frac{2^{-(i+j)/2}}{T}\{I + II\}.$$

By formula (3.67),

$$I \leq \sum_l (C_l + ML_l 2^{-l})(2^i A_{l,i} + 2^j A_{l,j}) \leq \sum_l (C_l + ML_l 2^{-l}) 2 \sum_i 2^i A_{l,i} \leq D_1.$$

As $\sum_i L_i 2^{-i} < \infty$, we must have $L_i \leq C2^i$ so $\sum_i L_i A_{i,j} \leq C$ again by (3.67). This and the Cauchy inequality give

$$II \leq 2M \left(\sum_l L_l A_{l,i}\right)^{1/2} \left(\sum_l L_l A_{l,j}\right)^{1/2} \leq D_2.$$

The bound for the remainder is therefore $O(2^{-(i+j)/2}T^{-1})$. For the main term, straightforward computation gives

$$\sum_{l,u} S_l\left(\frac{k}{T}\right) \sum_{s,t} \psi_{l,s-u}\psi_{j,k-s}\psi_{l,t-u}\psi_{i,k-t} = \sum_\tau c(k/T,\tau) \sum_n \psi_{i,n-\tau}\psi_{j,n},$$

which yields formula (3.73). Using (3.70) and (3.73) with $i = j$, we obtain

$$\begin{aligned}
\mathbb{E}\{\hat{c}(k/T,0)\} &= \sum_{j=-J(t)}^{-1} 2^j \left\{\sum_\tau c(k/T,\tau)\Psi_j(\tau) + O(2^{-j}/T)\right\} \\
&= \sum_\tau c(k/T,\tau)\delta_\tau - \sum_{j=-J(t)-1}^{-\infty} 2^j \sum_\tau c(k/T,\tau)\Psi_j(\tau) \\
&+ O(\log(T)/T) \\
&= c(k/T,0) + O\left(T^{-1}\sum_\tau c^*(\tau)\right) + O(\log(T)/T),
\end{aligned}$$

which proves the expectation by (3.43). For the variance, observe that, using Gaussianity, we have

$$\begin{aligned}\operatorname{cov}\left(I_{k,T}^{(i)}, I_{k,T}^{(j)}\right) &= 2\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n-\tau}\psi_{j,n} + O(2^{-(i+j)/2}T^{-1})\right)^2 \\ &= 2\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n-\tau}\psi_{j,n}\right)^2 + O(2^{-(i+j)/2}T^{-1}), \qquad (3.74)\end{aligned}$$

provided that (3.43) holds. Using (3.74), we finally obtain

$$\operatorname{Var}\left\{\hat{c}(k/T,0)\right\} = 2\sum_{i,j=-J(t)}^{-1} 2^{i+j}\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n-\tau}\psi_{j,n}\right)^2 + O(T^{-1}), \tag{3.75}$$

which completes the proof. □

Using Lemma 3.4.1, it is possible to show a similar result for $\tau \neq 0$. We show the derivation for the expectation of $\hat{c}(k/T,\tau)$ below. Using (3.73), we can write

$$\begin{aligned}&\mathbb{E}\left\{\hat{c}(k/T,\tau)\right\} = \\ &\quad \sum_{j=-J(t)}^{-1}\left(\sum_{l=-\infty}^{-1}(A^{-1})_{j,l}\Psi_l(\tau)\right)\left\{\sum_{n} c(k/T,n)\Psi_j(n) + O(2^{-j}/T)\right\} \\ &\quad \sum_{j=-\infty}^{-1}\sum_{l=-\infty}^{-1}(A^{-1})_{j,l}\Psi_l(\tau)\sum_{n} c(k/T,n)\Psi_j(n) + \\ &\quad -\sum_{j=-J(t)-1}^{-\infty}\sum_{l=-\infty}^{-1}(A^{-1})_{j,l}\Psi_l(\tau)\sum_{n} c(k/T,n) + \\ &\quad +O\left(T^{-1}\sum_{j=-J(t)}^{-1}2^{-j}\left|\sum_{l=-\infty}^{-1}(A^{-1})_{j,l}\Psi_l(\tau)\right|\right) = \\ &\quad I + II + III.\end{aligned}$$

We first concentrate on the main term.

$$\begin{aligned}
I &= \sum_{j=-\infty}^{-1} \sum_{l=-\infty}^{-1} (A^{-1})_{j,l} \Psi_l(\tau) \sum_n \sum_{i=-\infty}^{-1} S_i(k/T) \Psi_i(n) \Psi_j(n) \\
&= \sum_{i=-\infty}^{-1} \sum_{l=-\infty}^{-1} \left(\sum_{j=-\infty}^{-1} (A^{-1})_{j,l} A_{i,j} \right) S_i(k/T) \Psi_l(\tau) \\
&= \sum_{i=-\infty}^{-1} \sum_{l=-\infty}^{-1} \delta_{i-l} S_i(k/T) \Psi_l(\tau) \\
&= \sum_{i=-\infty}^{-1} S_i(k/T) \Psi_i(\tau) \\
&= c(k/T, \tau),
\end{aligned}$$

as expected. We now focus on the remainders. Using (3.69) and (3.43),

$$\begin{aligned}
|II| &= \left| \sum_{j=-J(t)-1}^{-\infty} \left(\sum_{l=-\infty}^{-1} (A^{-1})_{j,l} \Psi_l(\tau) \right) \sum_n c(k/T, n) \right| \\
&\leq \sum_{j=-J(t)-1}^{-\infty} \sum_{l=-\infty}^{-1} \left| (A^{-1})_{j,l} \right| \sum_n c^*(n) \\
&= O\left(\sum_{-\log(T)}^{-\infty} 2^{j/2} \right) \\
&= O(T^{-1/2}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
|III| &= O\left(T^{-1} \sum_{j=-J(t)}^{-1} 2^{-j} \left| \sum_{l=-\infty}^{-1} (A^{-1})_{j,l} \Psi_l(\tau) \right| \right) \\
&= O\left(T^{-1} \sum_{j=-J(t)}^{-1} 2^{-j/2} \right) \\
&= O(T^{-1/2}).
\end{aligned}$$

The above derivation leads to

$$\mathbb{E}\{\hat{c}(k/T, \tau)\} = c(k/T, \tau) + O(T^{-1/2}) \quad \text{for} \quad \tau \neq 0. \tag{3.76}$$

We conclude this section with a remark on the formula for the asymptotic variance (3.72). To gain some insight into this "complicated" formula, we examine how it simplifies in the case of a Gaussian white noise process $X_{t,T} = \sigma Z_t$, where $Z_t \sim N(0,1)$ i.i.d. Substituting $c(k/T, \tau) = \sigma^2 \delta_\tau$ and using $\sum_n \psi_{i,n}\psi_{j,n} = \delta_{i-j}$, we obtain

$$\begin{aligned}
\mathrm{Var}\left\{\hat{c}\left(\frac{k}{T}, 0\right)\right\} &= 2 \sum_{i,j=-J(t)}^{-1} 2^{i+j} \left(\sigma^2 \delta_{i-j}\right)^2 + O(T^{-1}) \\
&= 2\sigma^4 \sum_{i=-J(t)}^{-1} 2^{2i} + O(T^{-1}) \\
&\to 2/3\sigma^4.
\end{aligned}$$

Probably the simplest estimator of the local variance in the LSW model can be obtained by simply squaring the relevant observation:

$$\hat{c}_1\left(\frac{k}{T}, 0\right) = X_{k,T}^2. \tag{3.77}$$

Note that in the case $X_{t,T} = \sigma Z_t$, we have

$$\mathrm{Var}\left\{\hat{c}_1\left(\frac{k}{T}, 0\right)\right\} = 2\sigma^4 = 3\mathrm{Var}\left\{\hat{c}\left(\frac{k}{T}, 0\right)\right\}. \tag{3.78}$$

3.5 Prediction based on data

3.5.1 Nuisance parameters

Formula (3.72) shows the inconsistency of $\hat{c}(\cdot, \tau)$ for $\tau = 0$, and a similar result can be shown for $\tau \neq 0$. To obtain consistency, we smooth $\hat{c}(\cdot, \tau)$ for each fixed τ w.r.t. to the first argument. We use standard kernel smoothing with a bandwidth parameter g (later we propose a procedure for choosing g). For speed of computation, we use the same g for all τ. We denote the smoothed version of $\hat{c}$ by $\tilde{c}$.

In practice, the coefficients for the linear predictor are computed by inverting the system of linear equations (3.22), where the entries of $\mathbf{B}_{t,T}$ have been replaced by their empirical versions

$$\left(\hat{\mathbf{B}}_{t,T}\right)_{m,n} = \tilde{c}\left(\frac{m+n}{2T}, m-n\right). \tag{3.79}$$

The vector of the right-hand side coefficients in (3.22), given by

$$\left\{ \sum_{j=-1}^{-\infty} \sum_{k \in \mathbb{Z}} S_j \left(\frac{k}{T} \right) \psi_{j,k-(t+h-1)} \psi_{j,k-m} \right\}_{m=0,1,\dots,t-1} \tag{3.80}$$

can be approximated by the vector of corresponding local autocovariances

$$\left\{ c \left(\frac{t+h+m-1}{2T}, t+h-1-m \right) \right\}_{m=0,1,\dots,t-1}. \tag{3.81}$$

Therefore, we estimate it by

$$\left\{ \tilde{c} \left(\frac{t+h+m-1}{2T}, t+h-1-m \right) \right\}_{m=0,1,\dots,t-1}, \tag{3.82}$$

where $\tilde{c}(z, \tau)$ for $z > (t-1)/T$ has been extrapolated from the values of $\hat{c}(0, \tau), \dots, \hat{c}((t-1)/T, \tau)$ using the same kernel smoothing procedure with bandwidth g.

To achieve a greater forecast accuracy in practice, we reduce the dimension of the system of prediction equations (3.22) by considering a "clipped" predictor

$$\hat{X}^{(p)}_{t+h-1,T} = \sum_{s=t-p}^{t-1} b^{(h)}_{t-1-s,T} X_{s,T}, \tag{3.83}$$

where the index p needs to be chosen from the set $\{1, \dots, t\}$. The matrix of the resulting system of empirical prediction equations is now of dimension $p \times p$, instead of $(t-1) \times (t-1)$, and its entries are estimated using the procedure described above. The construction (3.83) is reminiscent of the classical idea of AR(p) approximation for stationary processes.

In order for the forecasting to be successful, the two nuisance parameters g and p need to be chosen in a data-driven manner. Section 3.5.3 describes a computational procedure for performing this selection.

3.5.2 Future observations in rescaled time

An important ingredient of the rescaled time concept, introduced in Section 2.2.4, is that the data come in the form of a triangular array whose rows correspond to different stochastic processes, only linked through the asymptotic wavelet spectrum sampled on a finer and finer grid. This mechanism is inherently different to what we observe in practice, where, typically, observations arrive one by one and neither the values of the "old" observations, nor their corresponding second-order structure, change when a new observation arrives.

One way to reconcile the practical setup with the theory is to assume that for an observed process $X_0, \ldots, X_{t-1}$, there exists a doubly-indexed LSW$_2$ process $\mathbf{Y}$ such that $X_k = Y_{k,T}$ for $k = 0, \ldots, t-1$. When a new observation X_t arrives, the underlying LSW$_2$ process changes, i.e. there exists another LSW$_2$ process $\mathbf{Z}$ such that $X_k = Z_{k,T+1}$ for $k = 0, \ldots, t$. An essential point underlying the adaptive algorithm of the next subsection is that the spectra of $\mathbf{Y}$ and $\mathbf{Z}$ are close to each other, due to the above construction and the regularity assumptions imposed by the definition of an LSW$_2$ process (in particular, the Lipschitz continuity of $W_j(z)$).

For clarity of presentation, we assume from now on that $h = 1$. The objective of the algorithm is to choose appropriate values of the two nuisance parameters g and p (see the next subsection) in order to forecast X_t from $X_0, \ldots, X_{t-1}$. Suppose that these parameters have been selected well, i.e. that the forecasting has been successful. The closeness of the two spectra implies that we can also expect to successfully forecast X_{t+1} from $X_0, \ldots, X_t$ using the same, or possibly "neighbouring", values of the nuisance parameters.

Bearing in mind the above discussion, we introduce the algorithm with a slight abuse of notation: we drop the second subscript when referring to the observed time series.

3.5.3 Data-driven choice of parameters

The idea of the procedure is to start with some initial values of p and g and to gradually update them using a criterion which measures how well the series gets predicted using a given pair of parameters. This type of approach is in the spirit of *adaptive forecasting* (Ledolter (1981)).

Suppose that we observe the series up to X_{t-1} and want to forecast X_t using an appropriately chosen pair (p, g). The idea of the method is to move backwards by s observations and choose an initial pair $\left(p_0^{(0)}, g_0^{(0)}\right)$ for predicting X_{t-s} from the observed series up to X_{t-s-1}. We compute the forecast of X_{t-s} using not only $\left(p_0^{(0)}, g_0^{(0)}\right)$, but also the 8 "neighbouring" parameter pairs

$$\left(p_0^{(0)} - 1, g_0^{(0)} - \eta\right), \left(p_0^{(0)}, g_0^{(0)} - \eta\right), \dots, \left(p_0^{(0)} + 1, g_0^{(0)} + \eta\right),$$

for a pre-selected constant η. As the true value of X_{t-s} is known, we are able to use a preset criterion to compare the 9 prediction results, and we set $\left(p_1^{(0)}, g_1^{(0)}\right)$ to be the pair that corresponds to the best result. In the next step, we use the pair $\left(p_1^{(0)}, g_1^{(0)}\right)$, as well as its 8 neighbours, to predict X_{t-s+1} from $X_0, \dots, X_{t-s}$, and obtain $\left(p_2^{(0)}, g_2^{(0)}\right)$ as the pair which resulted in the best forecast. Continuing in the same fashion until we reach X_{t-1}, we finally obtain an updated pair $\left(p_s^{(0)}, g_s^{(0)}\right)$, which is used to perform the actual prediction of X_t.

Several different criteria can be used to compare the quality of the pairs of parameters at each step. Denote by $\hat{X}_{t-i}(p, g)$ the predictor of X_{t-i} computed using the pair (p, g), and by $I_{t-i}(p, g)$ — the corresponding 95% *prediction interval* based on the assumption of Gaussianity:

$$I_{t-i}(p, g) = \left[-1.96\, \hat{\sigma}_{t-i}(p, g) + \hat{X}_{t-i}(p, g)\, , 1.96\, \hat{\sigma}_{t-i}(p, g) + \hat{X}_{t-i}(p, g)\right], \quad (3.84)$$

where $\hat{\sigma}^2_{t-i}(p, g)$ is the estimate of $\mathrm{MSPE}(\hat{X}_{t-i}(p, g), X_{t-i})$ computed using formula (3.12) with the remainder neglected. The criterion which we use in the simulations reported in the next section is to compute

$$\frac{\left|X_{t-i} - \hat{X}_{t-i}(p, g)\right|}{\mathrm{length}\{I_{t-i}(p, g)\}}$$

for each of the 9 pairs at each step of the procedure and select the updated pair as the one which minimises this ratio.

We also need to choose the initial parameters $\left(p_0^{(0)}, g_0^{(0)}\right)$ and the number s of data points at the end of the series which are used in the procedure. We suggest that s should be set to the length of the largest segment at the end of the series which does not contain any apparent breakpoints observed after a visual inspection. If, after a single pass along the segment $X_{t-s}, \ldots, X_{t-1}$, the forecasts are still inaccurate, then one or more further passes may be necessary: one possibility is then to set $\left(p_0^{(n)}, g_0^{(n)}\right) := \left(p_s^{(n-1)}, g_s^{(n-1)}\right)$ and proceed as before.

Note that the procedure is completely on-line: when the observation X_t becomes available, only a single update of the pair $\left(p_s^{(0)}, g_s^{(0)}\right)$ is needed to obtain a "good" pair of parameters for predicting X_{t+1}.

There are, obviously, many possible variants of the algorithm. Possible modifications include, for example, using a different criterion, restricting the allowed parameter space for (p, g), penalising certain regions of the parameter space, or permitting more than one parameter update at each time point.

The following section presents an application of the algorithm to a real data set. A more theoretical study of this algorithm is left for future work.

3.6 Application of the predictor to real data

In this section, we study the wind speed anomaly index, i.e. its standardised deviation from the mean, in a specific region of the Pacific (12-2N, 160E-70W). Modelling this anomaly helps to understand the El Niño effect in that region (see Philander (1990) for a detailed overview). The time series composed of $T = 910$ monthly observations is available free of charge at `http://tao.atmos.washington.edu/data_sets/eqpacmeridwindts`. Figure 3.1 shows the plot of the series.

Throughout this section, we use Haar wavelets to estimate the local (co)variance. Having provisionally made a safe assumption of the possible non-stationarity of the data, we first attempt to find a suitable pair of parameters (p, g) which will be used for forecasting the series. By inspecting the acf of the series, and by trying different values of the bandwidth, we have found that the pair $(7, 70)$ works well for many segments of the data; indeed, the segment of 100 observations from

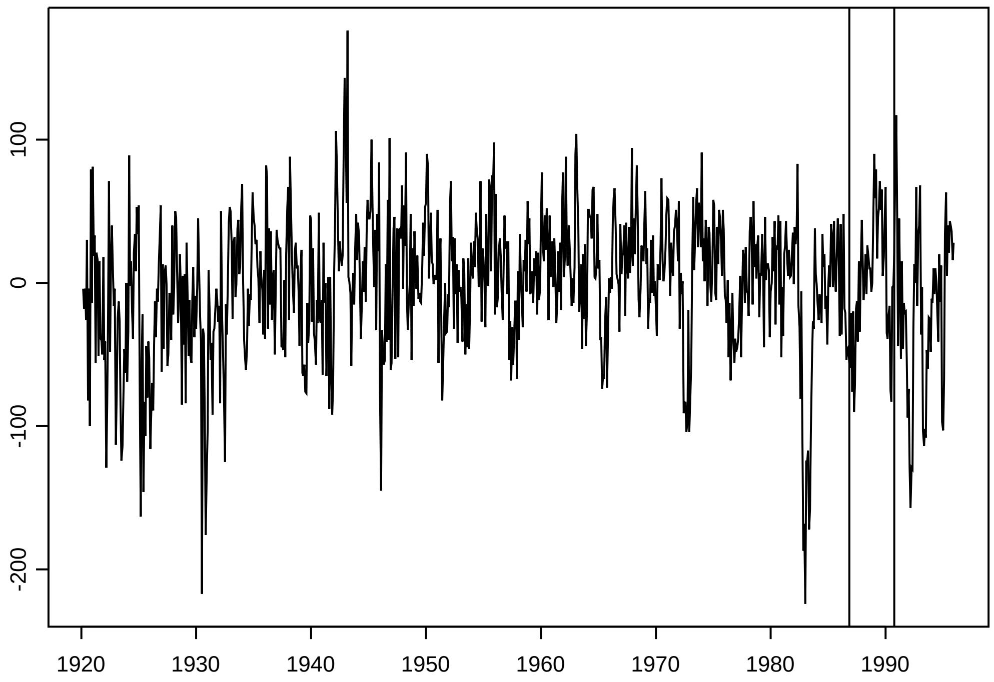

Figure 3.1: The wind anomaly index (in cm/s). The two vertical lines indicate the segment shown in Figure 3.2.

June 1928 to October 1936 gets predicted very accurately in one-step prediction: 96% of the actual observations are contained in the corresponding 95% prediction intervals (formula (3.84)).

However, the pair $(7, 70)$ does not appear to be uniformly well suited for forecasting the whole series. For example, in the segment of 40 observations between November 1986 and February 1990, only 5% of the observations fall into the corresponding one-step prediction intervals computed using the above pair of parameters. This provides strong evidence that the series is non-stationary (indeed, if it was stationary, we could expect to obtain a similar percentage of accurately predicted values in both segments).

Motivated by the above observation, we now apply the algorithm described in the previous section to the segment of 40 observations mentioned above, setting the

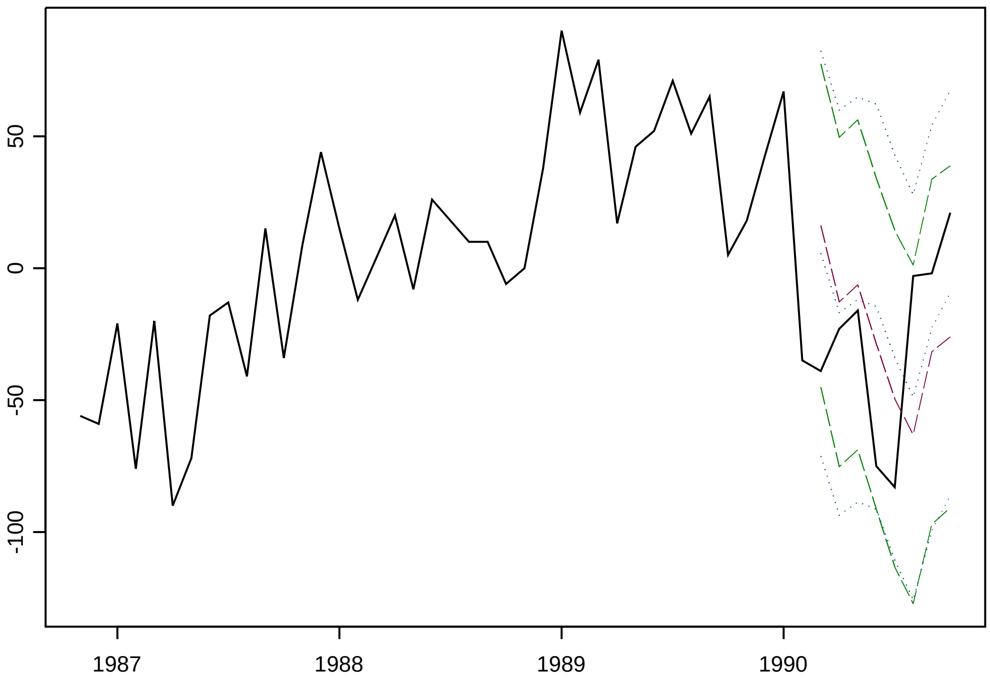

Figure 3.2: Comparison between the one-step prediction in the LSW_2 model (dashed lines) and AR (dotted lines). The middle line is the predicted value, the top (bottom) line is the upper (lower) end of the corresponding 95% prediction interval.

initial parameters to $(7, 70)$. After the first pass along the segment, the parameters drift up to $(14, 90)$, and 85% of the observations fall within the prediction intervals, which is indeed a dramatic improvement over the 5% obtained without applying the adaptive algorithm. In the second pass, we set the initial values to $(14, 90)$, and obtain a 92.5% coverage by the one-step prediction intervals, with the parameters drifting up to $(14, 104)$. In the last iteration, we finally obtain a 95% coverage, and the parameters get updated to $(14, 114)$. We now have every reason to believe that this pair of parameters is well suited for one-step prediction within a short distance of February 1990. Without performing any further updates, we apply the one-step forecasting procedure to predict, one by one, the eight observations which follow February 1990, the prediction parameters being fixed at $(14, 114)$. The results are plotted in Figure 3.2, which also compares our results to those obtained by means of AR modelling. At each time point, the order of the AR process is chosen as the one that minimises the AIC criterion, and then the parameters are estimated by means of the standard S-Plus routine. We observe that for both models, all of the true observed values fall within the corresponding one-step prediction intervals. However, the main gain obtained using our procedure is that the prediction intervals are on average 17.45% narrower in the case of our algorithm. This result is not peculiar to AR modelling as this percentage is also similar in comparison with other stationary models, like ARMA(2,10), believed to accurately fit the series. A similar phenomenon has also been observed at other points of the series.

3.7 Conclusion

In this chapter, we have investigated several theoretical and practical aspects of forecasting Gaussian LSW processes. As the model is Gaussian, we have considered the linear predictor where the coefficients minimise the Mean Square Prediction Error (MSPE). The exact MSPE, however, involves parameters which are unidentifiable in the LSW model. Therefore, we have considered the minimisation of an *approximation* to the MSPE which involves the (uniquely defined) wavelet spectrum. The derivation of our asymptotic results has been possible due to the rescaled time concept which is one of the ingredients of the LSW framework.

To overcome a theoretical difficulty arising in the approximation, we have introduced a slight modification to the LSW framework and called the new class the "LSW_2 model". All subsequent results have been derived for the new model. In particular, we have generalised Kolmogorov's formula for one-step-ahead prediction error.

In practice, the entries of the prediction matrix in the Yule-Walker equations need to be estimated. We have analysed the behaviour of the first two moments of the arising wavelet-based estimators and concluded that the estimators are asymptotically unbiased but inconsistent. Thus, the estimators need to be smoothed over time, and therefore we have to choose the smoothing parameter (e.g. the bandwidth of the smoothing kernel). Moreover, we need to reduce the dimension of the prediction equations to avoid too much inaccuracy of the resulting prediction coefficients due to estimation errors. We have proposed an automatic computational procedure for selecting these two parameters. Our algorithm is in the spirit of adaptive forecasting as it gradually updates the two parameters basing on the success of prediction.

We have applied our new algorithm to a time series of yearly values of the wind speed anomaly index in a specific region of the Pacific. Our non-parametric forecasting algorithm shows interesting advantages over the classical parametric alternative (AR forecasting). Moreover, we believe that one of the biggest advantages of our new algorithm is that it can be successfully applied to a variety of data sets, ranging from financial log-returns (Chapter 4) to series traditionally modelled as ARMA processes, including in particular data sets which are not, or do not appear to be, second-order stationary. The S-Plus routines implementing the algorithm, as well as the data set, are included on the associated CD.

Chapter 4

Modelling financial log-return series using wavelets and rescaled time

In this chapter, we attempt to model financial log-return series as locally stationary time series in a setup which combines wavelets and the rescaled time concept, and is closely related to the LSW framework of Nason *et al.* (2000). The initial motivation for this research can be summarised as follows:

1. As was mentioned in Section 2.2, stationary linear time series models cannot capture the "stylised facts" of financial log-return series and to preserve stationarity, non-linear models, such as (G)ARCH or Stochastic Volatility have been proposed. However, some authors (see the references in Section 2.2.2) have recently argued that even when non-linear models are used, non-stationary modelling may still be preferred. This provokes another interesting general question: once we abandon the assumption of stationarity, is non-linearity still needed to model financial log-returns accurately, or is it sufficient to stick to linear models?

2. Some authors observe that various economic factors operate at different time scales, which may translate into a possible "multiscale" mechanism underlying financial log-return series (see for example Calvet & Fisher (2001)). On

the other hand, wavelets are a commonly used tool in the analysis of multiscale phenomena, so a wavelet-based approach may prove to be a suitable modelling technique here.

3. The rescaled time idea, whereby the time-varying (first and) second order quantities of a process are defined on a compact interval, like in nonparametric regression, enables meaningful asymptotic considerations (see Section 2.2.4). Therefore the hope is that by modelling financial log-returns in a rescaled time framework we will be able to keep track of the asymptotic behaviour of the statistics of interest (e.g. sample autocorrelations of the squared returns), which may be helpful in explaining the commonly observed "stylised facts" of financial time series.

The chapter is organised as follows: in Section 4.1, we motivate our methodology by arguing that daily returns on the FTSE 100 index can be adequately modelled as Gaussian time-modulated white noise (TMWN). In Section 4.2, we introduce the LSW_3 model as a modification of the LSW framework of Nason *et al.* (2000), and show that Gaussian TMWN is a special case of an LSW_3 process. In Section 4.3, we provide theoretical evidence that LSW_3 processes can capture most of the stylised facts of financial time series modelling. In Section 4.4, we introduce a new (suitable for log-returns) estimation approach for LSW_3 processes, and demonstrate its superiority to the general method of Nason *et al.* (2000). In Section 4.5, we provide an interesting example of exploratory data analysis using the LSW_3 model. Finally, in Section 4.6, we apply the adaptive forecasting algorithm of Section 4.6 to log-returns, and provide a comparison with forecasts based on GARCH modelling.

4.1 Motivating example

In this section, we motivate our "linear non-stationary" approach by arguing that returns on the daily closing values of the FTSE 100 index can be adequately modelled as Gaussian time modulated white noise (TMWN), i.e. a process of the form $X_t = \sigma_t Z_t$, where σ_t is a deterministic sequence, and Z_t's are independent $N(0,1)$. In Section 4.2, we show that Gaussian TMWN is a special case of a wavelet-based time series model, closely related to the LSW framework recalled in Section 2.2.4.

For the purpose of this section, let X_t denote 2158 consecutive observations of logged and differenced daily closing values of the FTSE 100 index, from 22/23 October 1992 to 10/11 May 2001. The source of the data here, and throughout the rest of the chapter, is `http://bossa.pl/notowania/daneatech/metastock` (page in Polish).

X_t is plotted in the top left subfigure of Figure 4.1. Superimposed on the plot is an estimate $\hat{\sigma}_t$ of the local standard deviation σ_t (the estimate was obtained by smoothing X_t^2 using a Gaussian kernel with the bandwidth chosen by trial and error, and then square-rooting the result; see Section 4.4 for automatic methods of estimation). Following down the left-hand column, the next plot shows the sample autocorrelation of X_t, and the plot below it — the sample autocorrelation of X_t^2. The bottom left subfigure shows the Q-Q plot of X_t against the normal quantiles. From those plots, it is evident that X_t obeys the well-known "stylised facts": the sample autocorrelations of X_t are negligible, but the sample autocorrelations of X_t^2 are significant; volatility is clustered; the marginal distribution of X_t is heavy-tailed.

The right-hand column provides evidence that X_t can be modelled as Gaussian TMWN, which is a linear, but non-stationary stochastic process. Indeed, the top plot shows $Z_t = X_t / \hat{\sigma}_t$, and the plots in the 2nd and 3rd rows — the sample acf of Z_t and Z_t^2, respectively. The bottom right subfigure shows the Q-Q plot of Z_t against the normal quantiles. From the inspection of the sample autocorrelation functions of Z_t and Z_t^2, it appears that, as a first approximation, Z_t can be modelled fairly accurately as an i.i.d. sample of $N(0,1)$ variables. This in turn implies that X_t can be modelled as Gaussian TMWN: clearly, there exists a σ_t such that

$X_t = \sigma_t Z_t$ with Z_t i.i.d. $\sim N(0,1)$.

One of the consequences of the non-stationarity of X_t is the fact that the sample acf is simply not an appropriate tool for computing the acf of X_t or X_t^2. We would submit, and will argue this point later in the chapter, that the "long memory" effect in squared log-returns on indices is nothing else than a spurious effect of applying the sample acf to non-stationary data (see Mikosch & Starica (2003) for similar considerations in the GARCH framework).

Having demonstrated that daily FTSE 100 can be modelled as Gaussian TMWN, we now proceed to define our wavelet-based model (which is a modification of the LSW setup) and show that Gaussian TMWN is its special case. In Section 4.5, we come back to the example of FTSE 100 and model this series in our wavelet framework. We show that, in this way, more local features of the FTSE 100 data can be picked up.

4.2 Wavelet-based model

In the rescaled time framework, the TMWN process is defined as

$$X_{t,T} = \sigma\left(\frac{t}{T}\right) Z_{t,T}, \tag{4.1}$$

for $t = 0, \ldots, T-1$ and $T = 1, 2, \ldots$, where $\sigma \in C[0,1]$ is a "smooth" time-varying standard deviation function, and $Z_{t,T}$ are i.i.d. $N(0,1)$. Motivated by the discussion above, we wish to embed this process in a larger, "multiscale" stochastic framework. Clearly, the LSW (or LSW_2) model is a good candidate, possessing all the required characteristics: non-stationarity, linearity, rescaled time and a multiscale structure. However, we will now show that in the present formulation, the TMWN is not embedded in either the LSW or the LSW_2 model:

$$\text{TMWN} \not\subseteq \text{LSW}_{(2)}. \tag{4.2}$$

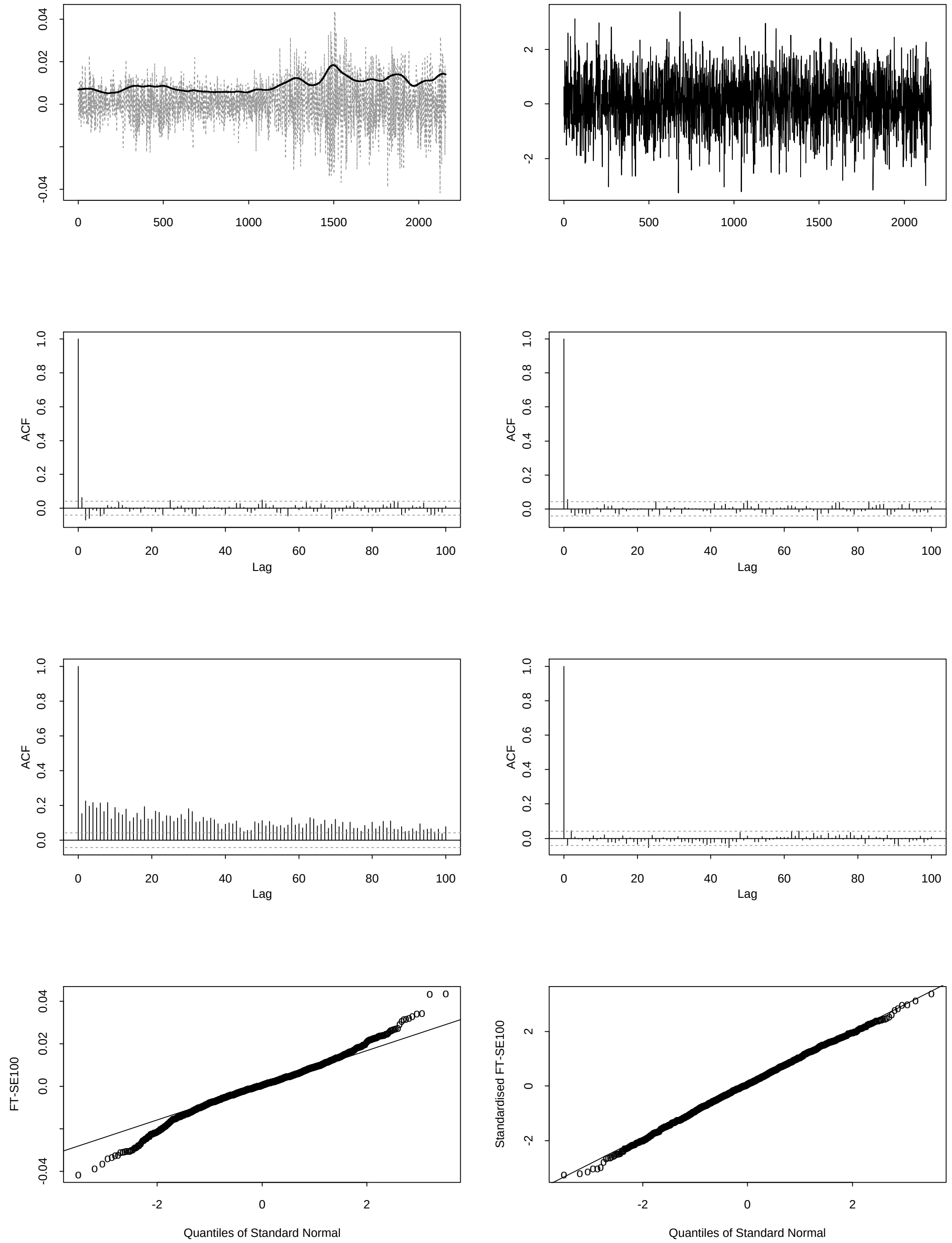

Figure 4.1: Left-hand column, from top to bottom: X_t with $\hat{\sigma}_t$ superimposed, acf of X_t, acf of X_t^2, qqnorm plot of X_t. Right-hand column, from top to bottom: Z_t, acf of Z_t, acf of Z_t^2, qqnorm plot of Z_t. See Section 4.1 for a discussion.

Indeed, multiplying both sides of (3.70) by $\sigma^2(z)$, we obtain that the wavelet spectrum of a TMWN process (4.1) is

$$S_j(z) = 2^j \sigma^2(z), \tag{4.3}$$

which leads to $W_j(z) = 2^{j/2}\sigma(z)$. Assume that $\sigma(z)$ is Lipschitz-continuous with the Lipschitz constant L. The Lipschitz constants L_j for W_j are therefore equal to $L_j = 2^{j/2}L$: however, this clearly violates the summability conditions (2.27) and (3.32).

To remedy this unwelcome situation, we introduce an alternative model which is a modified version of the LSW and LSW$_2$ setups. We call the new class "LSW$_3$".

Definition 4.2.1 *A triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$, for $T = 1, 2, \ldots$, is in the class of* LSW$_3$ *processes if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \omega_{j,k;T} \psi_{j,k}(t) \xi_{j,k}, \tag{4.4}$$

where $\psi_{j,k}(t)$ are nondecimated discrete wavelet vectors, $\omega_{j,k;T}$ are real constants, and $\{\xi_{j,k}\}_{j,k}$ are zero-mean orthonormal identically distributed random variables. Also, we assume that for each $j \le -1$, there exists a continuous function $W_j(z) : \mathbb{R} \to \mathbb{R}$ such that $S_j := W_j^2$ is Lipschitz with constant L_j and

- *$W_j(z) = W_j(0)$ for $z < 0$ and $W_j(z) = W_j(1)$ for $z > 1$,*
- *$\sum_{j=-\infty}^{-1} S_j < \infty$,*
- *the Lipschitz constants L_j satisfy*

$$\exists\, D < \infty \quad \forall j \le -1 \quad L_j 2^{-j} \le D, \tag{4.5}$$

- *there exists a sequence of constants C_j satisfying $\sum_{j=-\infty}^{-1} C_j < \infty$ such that, for each T,*

$$\sup_k |\omega_{j,k;T}^2 - S_j(k/T)| \le C_j/T \quad \forall j. \tag{4.6}$$

Three remarks are worth making at this point.

1. Nason *et al.* (2000) control the evolution of the second order structure of LSW processes by making certain assumptions on the Lipschitz constants of $\{W_j(z)\}_j$, as well as on the distances $\sup_k |\omega_{j,k;T} - W_j(k/T)|$ (we adopt the same convention in the definition of the LSW_2 process). However what is really required in the second-order theory is the Lipschitzness of $S_j(z) = W_j(z)^2$ and bounds on the distances $\sup_k |\omega^2_{j,k;T} - S_j(k/T)|$. Here is how the latter set of assumptions relates to the former: we have

$$\begin{aligned}
|\omega^2_{j,k;T} - S_j(k/T)| &= |\omega_{j,k;T} - W_j(k/T)||\omega_{j,k;T} - W_j(k/T) + 2W_j(k/T)| \\
&\leq C_j/T\,(C_j/T + 2W_j(z)) \\
&\leq C_j/T\,(C_j/T + 2\,(W_j(z) \vee 1)) \\
&\leq C_j/T\,(C_j/T + 2\,(S_j(z) \vee 1)) \\
&\leq C_j/T \left(C_j/T + 2 \left(\sup_z \sum_j S_j(z) \vee 1 \right) \right) \\
&\leq K_1 C_j/T.
\end{aligned}$$

Also,

$$\begin{aligned}
&|S_j((k+l)/T) - S_j(k/T)| \leq \\
&\quad L_j|l|/T|W_j((k+l)/T) - W_j(k/T) + 2W_j(k/T)| \leq \\
&\quad L_j|l|/T \left(L_j|l|/T + 2 \left(\sup_z \sum_j S_j(z) \vee 1 \right) \right) \leq \\
&\quad K_2 L_j |l|/T.
\end{aligned}$$

Therefore, in deriving rates of convergence in the LSW (LSW_2) model, we can "ignore" the constants K_1, K_2 and assume that $|\omega^2_{j,k;T} - S_j(k/T)| \leq C_j/T$ and $|S_j((k+l)/T) - S_j(k/T)| \leq L_j|l|/T$, which is implicitly done both in Nason *et al.* (2000) and in Chapter 3. Note that the definition of an LSW_3 process contains these two assumptions in an explicit form, but does not contain the "redundant" assumptions that $|\omega_{j,k;T} - W_j(k/T)| \leq C_j/T$ or $|W_j((k+l)/T) - W_j(k/T)| \leq L_j|l|/T$.

2. Note that the summability conditions (2.27) and (3.32) are stronger than (4.5). Indeed, $\sum_j 2^{-j} L_j < \infty$ implies (but is not equivalent to) $2^{-j} L_j \to 0$, which in turn implies (but is not equivalent to) (4.5).

3. Note that we also define $W_j(z)$ and $S_j(z)$ beyond the interval $[0,1]$ (we assume that W_j and S_j are continuous everywhere and constant outside $[0,1]$). This will be needed later in the proof of Proposition 4.2.1.

It is easy to see that Gaussian TMWN with σ Lipschitz satisfies the assumptions of Definition 4.2.1 with $\omega_{j,k;T} = W_j(k/T) = \sigma(k/T)2^{j/2}$.

The definitions of the local (co)variance and other quantities in the LSW$_3$ model are analogous to those in the LSW framework (see Section 2.2.4). Under the assumptions of Definition 4.2.1, the evolutionary wavelet spectrum $S_j(z)$ and the local autocovariance $c(z,\tau)$ remain uniquely defined (the proof of this statement is identical to Nason *et al.* (2000), Theorem 1). We are also in a position to prove the following proposition:

Proposition 4.2.1 *Under the assumptions of Definition 4.2.1,* $\|c_T - c\|_{L_\infty} = O(T^{-1}\log(T))$.

Proof. We have

$$\begin{aligned} |c_T(l/T,\tau) - c(l/T,\tau)| &= \left| \sum_{j,k} \left(\omega^2_{j,k;T} - S_j\left(\frac{l}{T}\right)\right) \psi_{j,k-l}\psi_{j,k-l-\tau} \right| \\ &\le \sum_{j,k} \left(\frac{C_j}{T} + \left| S_j\left(\frac{k}{T}\right) - S_j\left(\frac{l}{T}\right)\right| \right) |\psi_{j,k-l}\psi_{j,k-l-\tau}| \end{aligned} \tag{4.7}$$

Now, observe that on the one hand,

$$\left| S_j\left(\frac{k}{T}\right) - S_j\left(\frac{l}{T}\right)\right| \le \frac{L_j|k-l|}{T} \le \frac{L_j \mathcal{L}_j}{T}, \tag{4.8}$$

due to the compact support of ψ_j. However, on the other hand,

$$\left| S_j\left(\frac{k}{T}\right) - S_j\left(\frac{l}{T}\right)\right| \le L_j, \tag{4.9}$$

using the fact that S_j is Lipschitz (with constant L_j) and constant outside $[0,1]$. Therefore, continuing from (4.7), we have

$$\sum_{j,k} \left(\frac{C_j}{T} + \left| S_j\left(\frac{k}{T}\right) - S_j\left(\frac{l}{T}\right) \right| \right) |\psi_{j,k-l}\psi_{j,k-l-\tau}| \leq$$
$$O(T^{-1}) + \sum_j L_j \min\left(\frac{\mathcal{L}_j}{T}, 1\right), \tag{4.10}$$

by applying the Cauchy inequality for $\sum_k |\psi_{j,k-l}\psi_{j,k-l-\tau}|$. Let us now concentrate on the sum in (4.10). Recall that $\mathcal{L}_j \leq M2^{-j}$ for some constant M. We have

$$\begin{aligned}
\sum_j L_j \min\left(\frac{\mathcal{L}_j}{T}, 1\right) &\leq \sum_j L_j \min\left(\frac{M2^{-j}}{T}, 1\right) \\
&= \frac{M}{T} \sum_{j=-1}^{-\lfloor \log_2(T/M) \rfloor} L_j 2^{-j} + \sum_{j < -\lfloor \log_2(T/M) \rfloor} L_j \\
&\leq \frac{M}{T} \sum_{j=-1}^{-\lfloor \log_2(T) \rfloor} D + D \sum_{j < -\lfloor \log_2(T/M) \rfloor} 2^j \\
&\leq \frac{MD\log_2(T)}{T} + O(DM/T),
\end{aligned}$$

which completes the proof. □

Throughout the rest of the chapter, we will work with Definition 4.2.1, rather than Definition 2.2.1 or 3.2.2.

Innovations $\xi_{j,k}$. Throughout the chapter, we stick to $\xi_{j,k}$ i.i.d. $N(0,1)$. Gaussian innovations cannot account for skewed "marginal" distributions or extreme events, such as those present in the Nikkei index (left-hand plot in Figure 4.2) or the Dow Jones Industrial Average (DJIA) index (right-hand plot in Figure 4.2). We believe that these stylised facts can be captured by an appropriate choice of the distribution of $\xi_{j,k}$, e.g. a mixture of normals would have a better chance of picking up the occasional "spikes" in the series (see, again, Figure 4.2). Also, a combination of skewed innovations and "skewed" wavelets (i.e. such that $\sum_k \psi_{j,k}^3 \neq 0$) would be able to pick up the often-observed skewness of the log-return data. However, the emphasis in

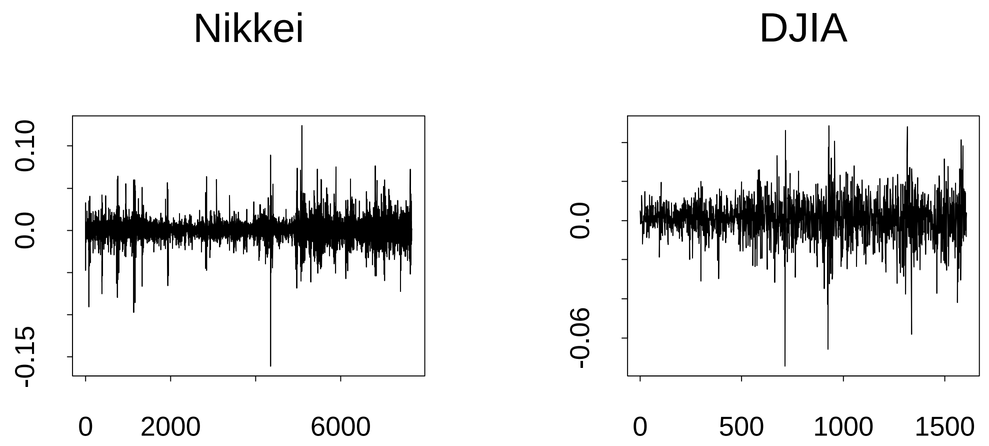

Figure 4.2: Left-hand plot: log-returns on daily closing values of Nikkei (5/6 Jan 1970 – 11/14 May 2001). Right-hand plot: log-returns on daily closing values of the Dow Jones Industrial Average (3/4 Jan 1995 – 10/11 May 2001).

this chapter is on the *non-stationarity* of the log-return series, and not on the possible *non-Gaussianity* of the innovations. Therefore, we restrict ourselves to Gaussian innovations in the theoretical considerations, leaving an extension to other distributions as an interesting direction for future study.

Trend. Throughout the chapter, we assume $\mathbb{E}(X_{t,T}) = 0$ (as is obvious from Definition 4.2.1). A more thorough study would also incorporate trend $\mu(t/T)$ in the model. This trend could then be estimated by wavelet methods, see e.g. von Sachs & MacGibbon (2000).

4.3 Explanation of the stylised facts

In this section, we demonstrate that Gaussian LSW_3 processes can successfully account for the following stylised facts of financial log-returns:

- heavy tails of the "marginal" distribution,
- negligible sample autocorrelations,
- non-negligible sample autocorrelations of the squares,

- clustering of volatility.

4.3.1 Heavy tails of the "marginal" distribution

In this section, we consider the sample second moment and the sample fourth moment:

$$\begin{aligned} m_2^T(\mathbf{X}) &= \frac{1}{T}\sum_{t=0}^{T-1} X_{t,T}^2 \\ m_4^T(\mathbf{X}) &= \frac{1}{T}\sum_{t=0}^{T-1} X_{t,T}^4. \end{aligned}$$

For stationary Gaussian processes, we could expect that $m_4^T(\mathbf{X})/(m_2^T(\mathbf{X}))^2 \sim 3$. However, the following demonstrates that this ratio is "spuriously" distorted if the variance $\sigma^2(z)$ of $X_{t,T}$ varies over time.

$$\begin{aligned} \mathbb{E}\left(m_4^T(\mathbf{X})\right) &= \frac{3}{T}\sum_{t=0}^{T-1}\sigma^4\left(\frac{t}{T}\right) + O(\log(T)/T) \\ &= \frac{3}{T}\sum_{t=0}^{T-1}\left(\sigma^2\left(\frac{t}{T}\right) - \frac{1}{T}\sum_{s=0}^{T-1}\sigma^2\left(\frac{s}{T}\right)\right)^2 \\ &+ 3\left(\frac{1}{T}\sum_{t=0}^{T-1}\sigma^2\left(\frac{t}{T}\right)\right)^2 + O(\log(T)/T) \\ &= \frac{3}{T}\sum_{t=0}^{T-1}\left(\sigma^2\left(\frac{t}{T}\right) - \frac{1}{T}\sum_{s=0}^{T-1}\sigma^2\left(\frac{s}{T}\right)\right)^2 \\ &+ 3\left(\mathbb{E}(m_2^T(\mathbf{X}))\right)^2 + O(\log(T)/T). \end{aligned}$$

For the purpose of this section, denote the first summand in the above formula by A^2. Obviously, $A^2 = 0$ for all T if and only if $\sigma^2(z)$ is constant w.r.t. z. Therefore, for a non-constant $\sigma^2(z)$ and for large T, we have

$$\frac{m_4^T(\mathbf{X})}{(m_2^T(\mathbf{X}))^2} \sim \frac{A^2}{(m_2^T(\mathbf{X}))^2} + 3 > 3. \tag{4.11}$$

The above relationship provides a heuristic explanation of the fact that the marginal distribution of processes with a non-constant variance appears heavy-tailed when the sample fourth moment and the sample second moment are (incorrectly) applied to them.

4.3.2 Sample autocorrelations of $X_{t,T}$ and $X_{t,T}^2$

Like in Mikosch & Starica (2003), we consider the sample autocovariance function

$$\gamma_T(\mathbf{X}, h) = \frac{1}{T} \sum_{t=0}^{T-1-h} X_{t,T} X_{t+h,T} - \left(\frac{1}{T} \sum_{t=0}^{T-1} X_{t,T} \right)^2, \tag{4.12}$$

and the sample autocorrelation function

$$\rho_T(\mathbf{X}, h) = \frac{\gamma_T(\mathbf{X}, h)}{\gamma_T(\mathbf{X}, 0)}. \tag{4.13}$$

Also, define the *scalogram*:

$$\tilde{S}_j^T = \frac{1}{T} \sum_{t=0}^{T-1} S_j \left(\frac{t}{T} \right). \tag{4.14}$$

The following proposition provides a representation of the expectation of the sample autocovariance in terms of the scalogram. The implications of the proposition are discussed after the proof.

Proposition 4.3.1 *Let an* LSW_3 *process* $X_{t,T}$ *satisfy*

$$\sup_z \sum_{\tau=-\infty}^{\infty} |c(z, \tau)| < \infty. \tag{4.15}$$

We have

$$\mathbb{E}(\gamma_T(\mathbf{X}, h)) = \sum_{j=-1}^{-\infty} \tilde{S}_j^T \Psi_j(h) + O\left(\frac{h + \log(T)}{T} \right). \tag{4.16}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E} \left\{ \frac{1}{T} \sum_{t=0}^{T-1-h} X_{t,T} X_{t+h,T} - \left(\frac{1}{T} \sum_{t=0}^{T-1} X_{t,T} \right)^2 \right\} = \\
&\frac{1}{T} \sum_{t=0}^{T-1-h} c\left(\frac{t}{T}, h \right) - \frac{1}{T^2} \sum_{t,t'=0}^{T-1} c\left(\frac{t}{T}, t - t' \right) + O\left(\frac{\log(T)}{T} \right) = \\
&\frac{1}{T} \sum_{t=0}^{T-1} c\left(\frac{t}{T}, h \right) - \frac{1}{T^2} \sum_{t,t'=0}^{T-1} c\left(\frac{t}{T}, t - t' \right) + O\left(\frac{h + \log(T)}{T} \right).
\end{aligned}$$

Let us now consider the second summand.

$$\begin{aligned}
\left| \frac{1}{T^2} \sum_{t,t'=0}^{T-1} c\left(\frac{t}{T}, t - t' \right) \right| &\leq \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{\tau=-\infty}^{\infty} \left| c\left(\frac{t}{T}, \tau \right) \right| \\
&\leq \frac{1}{T} \sup_z \sum_{\tau=-\infty}^{\infty} |c(z, \tau)| \\
&= O(T^{-1}),
\end{aligned}$$

by assumption (4.15). Noting that $\frac{1}{T}\sum_{t=0}^{T-1} c\left(\frac{t}{T}, h\right) = \sum_{j=-1}^{-\infty} \tilde{S}_j^T \Psi_j(h)$ completes the proof. $\square$

The representation (4.16) implies that the sample autocorrelations at positive lags will be negligible provided that $\sum_{j=-1}^{-\infty} \tilde{S}_j^T \Psi_j(h)$ is "close" to $C\delta_h$. By formula (3.70), this is guaranteed by $\tilde{S}_j^T$ being "close" to $C2^j$, which is indeed often the case, as the examples provided in Section 4.5 demonstrate. This would explain the frequently occurring negligible sample autocorrelations of log-returns.

An analogous proposition can be formulated for the sample autocovariance of $X_{t,T}^2$. The following result is true.

Proposition 4.3.2 *Let a Gaussian* LSW$_3$ *process $X_{t,T}$ satisfy (4.15). We have*

$$\begin{aligned}\mathbb{E}(\gamma_T(\mathbf{X^2}, h)) &= \frac{1}{T}\sum_{t=0}^{T-1}\left(\sigma^2\left(\frac{t}{T}\right) - \frac{1}{T}\sum_{s=0}^{T-1}\sigma^2\left(\frac{s}{T}\right)\right)^2 + \frac{2}{T}\sum_{t=0}^{T-1} c^2\left(\frac{t}{T}, h\right) \\ &+ O\left(\frac{h+\log(T)}{T}\right).\end{aligned}$$

Proof. Denote $\alpha(z) = \sum_\tau c^2(z,\tau)$. By assumption (4.15), $\sup_z \alpha(z) < \infty$. Using Gaussianity, we obtain

$$\begin{aligned}&\mathbb{E}\left\{\frac{1}{T}\sum_{t=0}^{T-1-h} X_{t,T}^2 X_{t+h,T}^2 - \left(\frac{1}{T}\sum_{t=0}^{T-1} X_{t,T}^2\right)^2\right\} = \\ &\frac{1}{T}\sum_{t=0}^{T-1-h}\left\{\sigma^2\left(\frac{t}{T}\right)\sigma^2\left(\frac{t+h}{T}\right) + 2\left(c\left(\frac{t}{T}, h\right)\right)^2\right\} + \\ &-\frac{1}{T^2}\sum_{t,t'=0}^{T-1}\left\{\sigma^2\left(\frac{t}{T}\right)\sigma^2\left(\frac{t'}{T}\right) + 2\left(c\left(\frac{t}{T}, t-t'\right)\right)^2\right\} + O\left(\frac{\log(T)}{T}\right) = \\ &\frac{1}{T}\sum_{t=0}^{T-1}\left\{\sigma^4\left(\frac{t}{T}\right) + 2\left(c\left(\frac{t}{T}, h\right)\right)^2\right\} + O\left(\frac{h}{T}\right) + \\ &-\left(\frac{1}{T}\sum_{t=0}^{T-1}\sigma^2\left(\frac{t}{T}\right)\right)^2 - \frac{2}{T^2}\sum_{t=0}^{T-1}\alpha\left(\frac{t}{T}\right) + O\left(\frac{\log(T)}{T}\right) = \\ &\frac{1}{T}\sum_{t=0}^{T-1}\left(\sigma^2\left(\frac{t}{T}\right) - \frac{1}{T}\sum_{s=0}^{T-1}\sigma^2\left(\frac{s}{T}\right)\right)^2 + \frac{2}{T}\sum_{t=0}^{T-1} c^2\left(\frac{t}{T}, h\right) + O\left(\frac{h+\log(T)}{T}\right),\end{aligned}$$

and the proof is completed. $\square$

For the purpose of this paragraph, denote the first summand of formula (4.17) by A^2, and the second one by $B^2(h)$. Two spurious effects can potentially be observed here. If the variance $\sigma^2(z)$ is non-constant, A^2 always gives a spurious positive contribution to the sample autocovariance. Note that A^2 is independent of h, which explains the fact that the sample autocovariance of the squares often decays very slowly (a feature which cannot be picked up by classical GARCH models, see again Mikosch & Starica (2003)). For extremely large h, the remainder $O(h/T)$ often makes the positive contribution of A^2 less pronounced.

The second spurious effect is due to $B^2(h)$, which distorts the information about the local autocovariance by averaging it over time. Things are not rectified in the case of the sample autocorrelation, either: as an example, consider again TMWN. For a non-constant $\sigma^2(z)$ and $h \neq 0$, we have

$$\rho_T(\mathbf{X^2}, h) = \frac{\gamma_T(\mathbf{X^2}, h)}{\gamma_T(\mathbf{X^2}, 0)} \sim \frac{A^2 + 0}{A^2 + B^2(0)} > \frac{0}{B^2(0)} = 0,$$

while, obviously, we would expect a good estimate to return a value close to zero. A similar mechanism works in the case of absolute values.

4.3.3 Clustering of volatility

The "clustering of volatility" or, in other words, a "slowly varying local variance" is indeed one of the features of LSW (LSW_2, LSW_3) modelling. Occasional "spikes" in the log-return series, see for example Figure 4.2, are clearly against this principle. Yet, we believe that this problem can be rectified by resorting to non-Gaussian innovations $\xi_{j,k}$, e.g. modelled by mixtures of normal variables. As mentioned in the penultimate paragraph of Section 4.2, a more thorough investigation of this possibility is left for future study.

4.4 Estimation

To estimate the spectrum, Nason *et al.* (2000) use the wavelet periodogram $I_{t,T}^{(j)}$, defined by formula (2.36). In our altered setup of Definition 4.2.1, we will also use the statistic defined by (2.36). The following proposition holds.

Proposition 4.4.1 *Let $X_{t,T}$ satisfy Definition 4.2.1. We have*

$$\mathbb{E}I_{t,T}^{(j)} = \sum_{i=-\infty}^{-1} S_i\left(\frac{t}{T}\right) A_{i,j} + O\left(\frac{2^{-j}\log(T)}{T}\right). \tag{4.17}$$

where A is defined by (2.34). In addition, if $X_{t,T}$ is Gaussian, then

$$\operatorname{Var}\left(I_{t,T}^{(j)}\right) = 2\left(\sum_i S_i\left(\frac{t}{T}\right) A_{i,j}\right)^2 + O\left(\frac{2^{-j}\log(T)}{T}\right). \tag{4.18}$$

Proof. In the proof, we use the orthonormality of $\xi_{j,k}$, the fact that $\mathcal{L}_j \leq M2^{-j}$ and formula (3.67). First note that

$$\sum_i \frac{C_i A_{i,j}}{T} \leq \frac{2^{-j}}{T}\sum_i C_i\left(\sum_j 2^j A_{i,j}\right) = \frac{2^{-j}}{T}\sum_i C_i = O\left(\frac{2^{-j}}{T}\right). \tag{4.19}$$

We have

$$\begin{aligned}
&\left|\mathbb{E}I_{p,T}^{(j)} - \sum_{i=-\infty}^{-1} S_i\left(\frac{p}{T}\right) A_{i,j}\right| \leq \\
&\quad \sum_{i=-1}^{-\infty}\sum_{k\in\mathbb{Z}}\left|\omega_{i,k;T}^2 - S_i\left(\frac{p}{T}\right)\right|\left(\sum_t \psi_{i,k-t}\psi_{j,p-t}\right)^2 \leq \\
&\quad \sum_{i=-1}^{-\infty}\sum_{k\in\mathbb{Z}}\left\{\left|S_i\left(\frac{k}{T}\right) - S_i\left(\frac{p}{T}\right)\right| + \frac{C_i}{T}\right\}\left(\sum_t \psi_{i,k-t}\psi_{j,p-t}\right)^2 \leq \\
&\quad \sum_i\left\{L_i \min\left(\frac{M\max(2^{-i},2^{-j})}{T},1\right) + \frac{C_i}{T}\right\} A_{i,j} \leq \\
&\quad \frac{M2^{-j}}{T}\left(\sum_{i=-1}^{j} L_i 2^{-i}2^{i}A_{i,j} + \sum_{i=j-1}^{-\lfloor\log_2(T/M)\rfloor} L_i 2^{-i}2^{j}A_{i,j}\right) + \\
&\quad + \sum_{i<-\lfloor\log_2(T/M)\rfloor} L_i + O(2^{-j}/T) \leq \\
&\quad \frac{M2^{-j}}{T}\sum_{i=-1}^{-\lfloor\log_2(T/M)\rfloor} L_i 2^{-i} + O(T^{-1}) + O(2^{-j}/T) = \\
&\quad O\left(\frac{2^{-j}\log(T)}{T}\right),
\end{aligned}$$

which proves the expectation. For the variance, first observe that

$$\begin{aligned}
\left|\sum_i S_i(z)A_{i,j}\right| &= \left|\sum_i S_i(z)\sum_\tau \Psi_i(\tau)\Psi_j(\tau)\right| \\
&= \left|\sum_\tau c(z,\tau)\Psi_j(\tau)\right| \\
&\leq \sup_z \sum_\tau |c(z,\tau)| \\
&< \infty
\end{aligned} \tag{4.20}$$

by assumption (4.15). Using Gaussianity, we have

$$\begin{aligned}
\mathrm{Var}\left(I_{p,T}^{(j)}\right) &= 2\left(\mathbb{E}\left(I_{p,T}^{(j)}\right)\right)^2 \\
&= 2\left(\sum_{i=-\infty}^{-1} S_i\left(\frac{p}{T}\right)A_{i,j} + O\left(\frac{2^{-j}\log(T)}{T}\right)\right)^2 \\
&= 2\left(\sum_{i=-\infty}^{-1} S_i\left(\frac{p}{T}\right)A_{i,j}\right)^2 + O\left(\frac{2^{-j}\log(T)}{T}\right),
\end{aligned}$$

where the last step uses (4.20). This completes the proof. □

The form of the remainder in (4.17) suggests that the estimator is more accurate for finer scales. However, like in Nason *et al.* (2000) and in Section 3.4, we normally compute the wavelet periodogram down to scale $-J(T)$, with J defined in Definition 2.2.1.

Formula (4.17) suggests the following method of estimating the spectrum: for each $t = 0, \ldots, T-1$, we solve the system of equations

$$I_{t,T}^{(j)} = \sum_i \hat{S}_i(t/T)A_{i,j}, \quad i,j = -1,\ldots,-J(T) \tag{4.21}$$

to obtain an approximately unbiased estimator $\hat{S}_j(t/T)$ of the spectrum $S_j(t/T)$ (see Nason *et al.* (2000) for details of this procedure in the LSW model).

However, formula (4.18) shows that the wavelet periodogram is not a consistent estimator and needs to be smoothed to obtain consistency. We can either first solve (4.21), and then smooth $\hat{S}_j(t/T)$, or first smooth $I_{t,T}^{(j)}$, and then solve (4.21). Following Nason *et al.* (2000), we prefer the latter option, as it is often easier to work out the distributional properties of $I_{t,T}^{(j)}$ than those of $\hat{S}_j(t/T)$, and therefore it is easier to justify the choice of smoothing parameters for $I_{t,T}^{(j)}$.

Smoothing the wavelet periodogram is by no means an easy task, due to an extremely low signal-to-noise ratio (for Gaussian series, neglecting the remainders, we have $\mathbb{E}\left(I_{t,T}^{(j)}\right) / \left\{\mathrm{Var}\left(I_{t,T}^{(j)}\right)\right\}^{1/2} \approx 1/\sqrt{2}$), and also to a significant amount of autocorrelation present in $I_{t,T}^{(j)}$. Nason *et al.* (2000) propose an adaptive wavelet denoising method whose performance will be discussed in Section 4.4.4.

In Section 4.4.1, we propose an alternative general methodology for smoothing the wavelet periodogram. Section 4.4.2 looks at two specific methods of smoothing, and Section 4.4.3 deals with inverting (4.21) in an approximate manner to ensure the nonnegativity of the estimated spectrum.

4.4.1 Generic algorithm

The approach which we propose here is based on the following observation. Denote by $\{d_{t,T}^{(j)}\}_{t=0}^{T-1}$ the sequence of non-decimated wavelet coefficients of $X_{t,T}$ at scale j (so that $I_{t,T}^{(j)} = \left(d_{t,T}^{(j)}\right)^2$). Often, financial log-returns exhibit little serial correlation (e.g. see the example in Section 4.1), so, by orthogonality of the *decimated* wavelets, the sequence

$$d_{0,T}^{(-1)}, d_{2,T}^{(-1)}, d_{4,T}^{(-1)}, \ldots, d_{T-2,T}^{(-1)}$$

as well as the sequence

$$d_{1,T}^{(-1)}, d_{3,T}^{(-1)}, d_{5,T}^{(-1)}, \ldots, d_{T-1,T}^{(-1)}$$

are each sequences of approximately uncorrelated random variables. At scale j, the same phenomenon is observed for sequences

$$d_{i,T}^{(j)}, d_{i+2^{-j},T}^{(j)}, \ldots, d_{i+T-2^{-j},T}^{(j)}, \quad i = 0, 1, \ldots, 2^{-j} - 1.$$

However, even if the original series $X_{t,T}$ exhibits some form of autocorrelation, the decimated sequences of wavelet coefficients will often be much less correlated.

This is the well-known "whitening" property of wavelets, see e.g. Vidakovic (1999), Section 9.5.3.

If $X_{t,T}$ is Gaussian, the lack of serial correlation in the decimated sequences also means lack of dependence, which in turn implies that the corresponding decimated subsequences of the wavelet periodogram

$$I^{(j)}_{i,T}, I^{(j)}_{i+2^{-j},T}, \dots, I^{(j)}_{i+T-2^{-j},T}, \quad i = 0, 1, \dots, 2^{-j} - 1. \tag{4.22}$$

are simply sequences of independent (gamma-distributed) random variables.

The above argument can only be made formal if $X_{t,T}$ is Gaussian TMWN. This is obviously a simplifying assumption, as clearly not every log-return sequence can be adequately modelled as such. However, it turns out that in practice, the assumption of the lack of dependence in the decimated subsequences of the wavelet periodogram leads to estimators which perform well numerically (on simulated data) and are visually appealing (on both simulated and real data). In other words, the departure from the TMWN setting often turns out not to be significant enough to prevent us from treating the decimated subsequences of $I^{(j)}_{t,T}$ as independent.

Having made the assumption of independence, we now proceed as follows:

1. Fix j.

2. For $i = 0, 1, \dots, 2^{-j} - 1$, pick the decimated sequence

$$I^{(j)}_{i,T}, I^{(j)}_{i+2^{-j},T}, \dots, I^{(j)}_{i+T-2^{-j},T}$$

   and smooth it using a preselected method, *with the smoothing parameter(s) chosen by cross-validation (CV)*. CV stands a chance of performing well here, due to the lack of dependence between the variables. For example, the technique of Ombao *et al.* (2001b) can be applied, as we are also dealing with a sample of independent gamma variates, like in periodogram smoothing. In Section 4.4.2, we explore two other methods in which the smoothing parameter is chosen by CV.

3. Interpolate the smoothed sequence at all the points $0, 1, \dots, T-1$ (e.g. using linear interpolation). Denote the interpolated smoothed sequence by
$$\left\{\tilde{I}^{(i,j)}_{t,T}\right\}_{t=0}^{T-1}.$$

4. Finally, compute the estimate of the wavelet periodogram as the average of the estimates $\tilde{I}^{(i,j)}_{\cdot,T}$, for $i = 0, 1, \dots, 2^{-j} - 1$:
$$\hat{I}^{(j)}_{t,T} = \sum_{i=0}^{2^{-j}-1} \tilde{I}^{(i,j)}_{t,T}.$$

For coarser scales, where it is not possible to smooth the decimated sequences accurately as they are too short, we estimate $\hat{I}^{(j)}_{t,T}$ by a constant: $\hat{I}^{(j)}_{t,T} = 1/T \sum_{l=0}^{T-1} I^{(j)}_{l,T}$.

The estimates $\hat{I}^{(j)}_{t,T}$ can now be substituted into the systems of linear equations
$$\hat{I}^{(j)}_{t,T} = \sum_i \hat{S}_i(t/T) A_{i,j}. \tag{4.23}$$

CV for dependent data. CV "as it is" does not perform well when the errors are dependent and some methods for correcting CV to this setting have been developed, see for example Altman (1990). However, they all work for stationary noise and require an estimate of the autocovariance. In our setting, finding such an estimate implies finding a pre-estimate of the signal itself. To avoid this nuisance, we prefer to work with independent decimated subsequences.

4.4.2 Smoothing the decimated periodogram

In step 2. of the algorithm of Section 4.4.1, we apply a smoothing procedure to the decimated subsequences of the wavelet periodogram. In this section, we consider the use of two smoothing methods: *cubic B-splines* (see Hastie & Tibshirani (1990) for details) and *translation-invariant nonlinear wavelet smoothing* (see Nason *et al.* (2000)).

The benefits of using cubic B-splines are the following.

- The method performs well (see Section 4.4.4).

- Most statistical packages provide a fast implementation of this method. For example, we use the S-Plus routine `smooth.spline`, which automatically selects the smoothing parameter by cross-validation.

- Numerical examples suggest that the method is fairly robust to the misspecification of the local variance of the noise. This feature is particularly attractive: in our setting, the variance of the noise depends on the signal (see formulas (4.17) and (4.18)), and, therefore, an accurate estimate of the variance would require an accurate estimate of the signal. In practice, it seems sufficient to supply constant variance to `smooth.spline`, see the results in Section 4.4.4.

The advantages of using translation-invariant nonlinear wavelet smoothing are as follows.

- The method performs well (see Section 4.4.4).

- The only smoothing parameter to be chosen is the "primary resolution", above which universal thresholding is applied with the threshold (suitable for chi-squares) as in Nason *et al.* (2000). For speed of computation, we do not choose the threshold by cross-validation, even though in theory this could also be considered. There are only $\log_2(T)$ primary resolution levels to choose from, which makes the choice potentially easier than, say, the choice of bandwidth in kernel smoothing. We perform the selection by "leave-half-out" cross-validation like in Nason (1996). The accurate choice of the primary resolution is extremely important in this context, as the numerical example of Section 4.4.4 powerfully demonstrates.

- Unlike linear methods, this nonlinear technique is capable of detecting abrupt changes in the wavelet periodogram.

4.4.3 Estimating the spectrum with guaranteed nonnegativity

The evolutionary wavelet spectrum $S_j(z)$ is a nonnegative quantity so it would also be desirable if $\hat{S}_j(k/T)$ was guaranteed to be nonnegative. This can be achieved, for example, by replacing the system of equations (4.23) by a Linear Complementarity Problem (LCP; see e.g. Murty (1988)):

$$\begin{aligned} A\hat{\mathbf{S}}(k/T) &\geq \hat{\mathbf{I}}_{k,T} \\ \hat{\mathbf{S}}(k/T) &\geq 0 \\ \left(A\hat{\mathbf{S}}(k/T) - \hat{\mathbf{I}}_{k,T}\right)\hat{\mathbf{S}}(k/T) &= 0. \end{aligned}$$

The above LCP can be solved using e.g. successive over-relaxation.

Let $\hat{S}_j^{\mathrm{LCP}}(k/T)$ denote the estimate of $S_j(k/T)$ obtained using the LCP formulation, and $\hat{S}_j^{\mathrm{INV}}(k/T)$ — using the simple inversion of formula (4.23). By (2.33), we estimate the local variance $\sigma^2(k/T)$ in each case by

$$\begin{aligned} \hat{\sigma}^2(k/T)^{(\mathrm{LCP})} &= \sum_{j=-J(T)}^{-1} \hat{S}_j^{\mathrm{LCP}}(k/T) \\ \hat{\sigma}^2(k/T)^{(\mathrm{INV})} &= \sum_{j=-J(T)}^{-1} \hat{S}_j^{\mathrm{INV}}(k/T). \end{aligned}$$

In practice, $\hat{\sigma}^2(k/T)^{(\mathrm{INV})}$ is often a much more accurate estimator of the local variance. In order to combine this feature with the guaranteed nonnegativity of the spectrum, we rescale the LCP-based estimator to yield the final estimators of $S_j(k/T)$ and $\sigma^2(k/T)$:

$$\hat{S}_j(k/T) = \hat{\sigma}^2(k/T)^{(\mathrm{INV})} \frac{\hat{S}_j^{\mathrm{LCP}}(k/T)}{\hat{\sigma}^2(k/T)^{(\mathrm{LCP})}} \tag{4.24}$$

$$\hat{\sigma}^2(k/T) = \sum_{j=-J(T)}^{-1} \hat{S}_j(k/T). \tag{4.25}$$

As explained in Sections 4.4.1 and 4.4.2, $\hat{S}_j(k/T)$ depends on the method used for smoothing the wavelet periodogram. The next section briefly compares the performance of the estimators based on cubic B-splines and on nonlinear wavelet denoising.

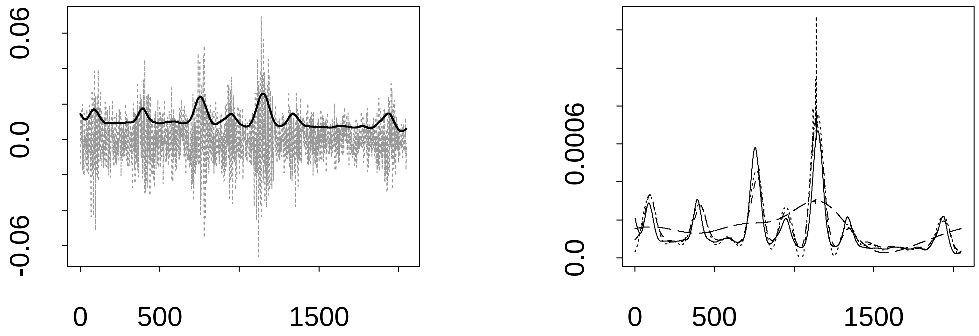

Figure 4.3: Left-hand plot: sample path from Gaussian TMWN model with time-varying standard deviation superimposed. Right-hand plot: time-varying variance (solid), its estimate using splines (dot-dashed), its estimate using nonlinear wavelet thresholding (dotted), and its estimate using nonlinear wavelet thresholding with default parameters (dashed).

4.4.4 Numerical results

The left-hand plot in Figure 4.3 shows a sample path from the Gaussian TMWN model with the superimposed contrived time-varying standard deviation. We estimate the time-varying local variance (the square of the time-varying standard deviation) by adding up estimators of the Haar wavelet spectrum over scales (see formula (4.25)). The right-hand plot shows

- the time-varying variance (solid line);
- an estimate obtained using spline smoothing with the smoothing parameter chosen by cross-validation (dot-dashed line);
- an estimate obtained using translation-invariant nonlinear wavelet smoothing with Daubechies' least asymmetric wavelet with 10 vanishing moments where the primary resolution was chosen by cross-validation (dotted line);

	default	splines	wavelets
mean of d_{σ^2}	1197	189	195
mean of d_S	464	243	276

Table 4.1: Values of the criterion functions averaged over 25 simulations. "Default" is the method of Nason *et al.* (2000) with default parameters, "splines" is our method using spline smoothing and "wavelets" is our method using translation-invariant nonlinear wavelet smoothing.

- an estimate obtained using the same wavelet method (dashed line) but with default parameters except the `smooth.dev` parameter in the `ewspec` routine (Nason (1998)) was set to `var` as recommended by G.P. Nason (personal communication).

While the two estimates with the smoothing parameter chosen by cross-validation almost coincide with each other and with the true time-varying variance (except for the spurious spike in the wavelet estimate), the default estimate by Nason *et al.* (2000) oversmooths. This is due to the fact that the primary resolution (PR) in the latter method is not chosen in a data-driven way but instead a fixed PR is used.

For the same Gaussian TMWN process, we assessed the performance of the three methods discussed above basing on 25 simulated sample paths. We used two criterion functions — one for the Haar spectrum:

$$d_S(\hat{S}, S) = \left[\frac{10^{11}}{T} \sum_{i=-J(T)}^{-1} \sum_{t=0}^{T-1} \left(\hat{S}_i\left(\frac{t}{T}\right) - S_i\left(\frac{t}{T}\right) \right)^2 \right], \tag{4.26}$$

and the other for the variance:

$$d_{\sigma^2}(\hat{\sigma}^2, \sigma^2) = \left[\frac{10^{11}}{T} \sum_{t=0}^{T-1} \left(\hat{\sigma}^2\left(\frac{t}{T}\right) - \sigma^2\left(\frac{t}{T}\right) \right)^2 \right]. \tag{4.27}$$

The values in Table 4.1 confirm our earlier observation that the two estimators in which the choice of the smoothing parameter is performed by cross-validation give very similar results.

4.5 Exploratory data analysis

In this section, we look at two examples of data analysis using the LSW_3 methodology (the examples are related to each other). The first one uses the Haar scalogram (see formula (4.14)), and the other — the full evolutionary Haar wavelet spectrum.

4.5.1 Analysis based on the scalogram

In this subsection, we compute the Haar scalogram for four series:

- $X_{t,T}$: the last 1024 observations of the artificial simulated Gaussian TMWN of Figure 4.3,
- $F_{t,T}$: the last 1024 observations of the FTSE 100 series of Figure 4.1,
- $N_{t,T}$: the last 1024 observations of the Nikkei series of Figure 4.2,
- $D_{t,T}$: the last 1024 observations of the Dow Jones IA series of Figure 4.2.

Figure 4.4 shows logged scalograms for $X_{t,T}$, $F_{t,T}$, $N_{t,T}$ and $D_{t,T}$ (solid lines), plotted against $-j = 1, 2, \ldots, 10$. Dotted lines are theoretical log-scalograms of corresponding time-modulated white noise processes with the same time-varying variances. As $X_{t,T}$ actually *is* a time-modulated white noise process, and its log-scalogram is substantially deviated from the corresponding dotted straight line for scales $-6, -7, \ldots, -10$, and slightly deviated for scales $-4, -5$, we suspect that for a series of length 1024, the scalogram is a relatively reliable estimator for scales $-1, -2, \ldots, -5$ (hence the vertical line at $-j = 5$), and a very reliable one for scales $-1, -2, -3$ (hence the vertical line at $-j = 3$).

Looking at the 3 finest scales ($-j = 1, 2, 3$), it seems that Dow Jones and Nikkei are reasonably close to TMWN. However, FTSE 100, which was provisionally modelled as Gaussian TMWN in Section 4.1, shows a substantial deviation from this setting, especially at scale $j = -2$, where the mean spectrum is clearly greater than what it should be if FTSE 100 were to be close to TMWN. Indeed, to assess

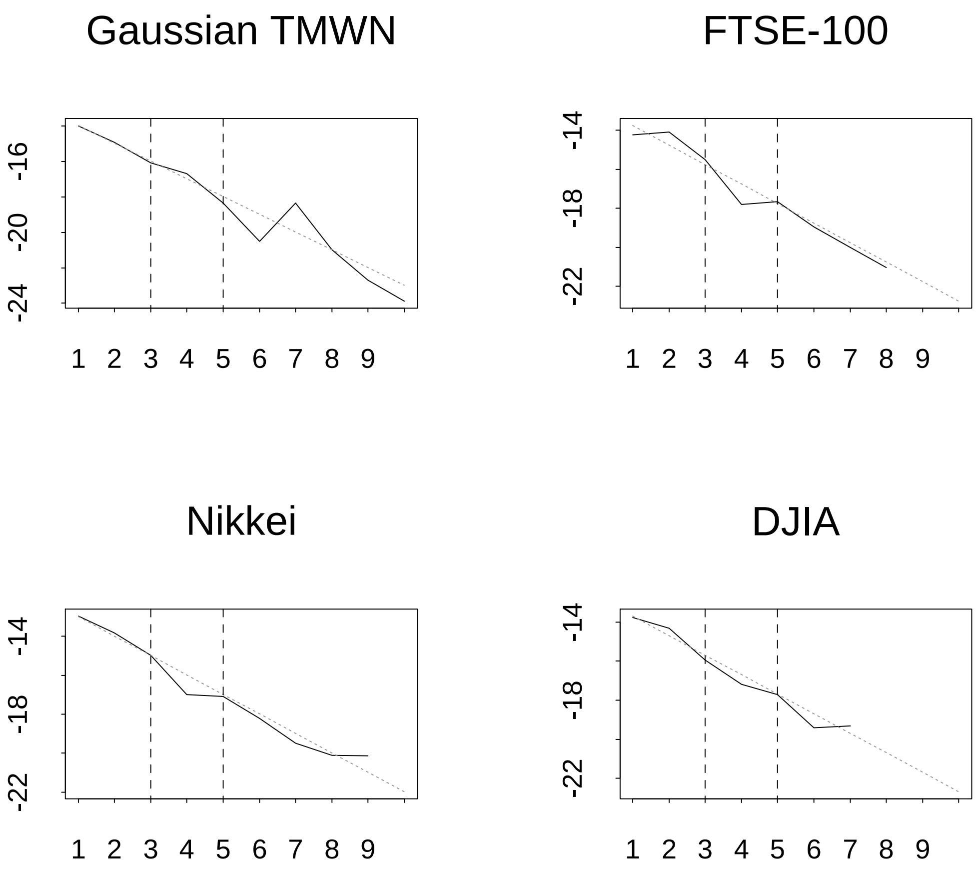

Figure 4.4: Solid lines: log-scalograms of $X_{t,T}$ (top left), $F_{t,T}$ (top right), $N_{t,T}$ (bottom left) and $D_{t,T}$ (bottom right), plotted against $-j$. Dotted lines: theoretical scalograms if the processes were (time-modulated) white noise (not necessarily Gaussian). Dashed lines: $-j = 3, 5$ (see text for discussion).

the validity of this statement, we have simulated 1000 independent sample paths of the standard white noise, and computed the Haar scalogram for each of them. In each case, the empirical scalogram for $j = -1$ was larger than that for $j = -2$, unlike the FTSE 100 case. The outcome of this experiment seems to confirm our initial judgement that the deviation of FTSE 100 from the TMWN setting is significant.

By formula (4.16), a large scalogram at scale $j = -2$ implies a significant contribution of the summand $\tilde{S}_{-2}^{T}\Psi_{-2}(h)$ to the sample autocovariance. For Haar wavelets, $\Psi_{-2}(\cdot)$ is supported on $h = -3, \ldots, 3$, and is plotted in the left plot of Figure 4.5. It is positive for $h = \pm 1$ and negative for $h = \pm 2, \pm 3$. Therefore, if the contribution of the spectrum at scale $j = -2$ is significant enough, we can expect that the sample autocorrelation of $F_{t,T}$ will be significant positive for $h = 1$, and significant negative for $h = 2, 3$. The right-hand plot in Figure 4.5 shows that this is indeed the case. The shape of the acf function of $F_{t,T}$ is very similar to the structure of Ψ_{-2}.

Figure 4.1 shows that the same pattern is present in the sample autocorrelation of the whole FTSE 100 series, and not only in $F_{t,T}$ (= the last 1024 observations of FTSE 100). However, the pattern is much less visible in the sample autocorrelation of the standardised FTSE 100 (series Z_t in Figure 4.1). This may suggest, for example, that this autocorrelation structure (positive dependence at lag 1, negative at lags 2 and 3), may be present in a stretch of high volatility, which has a significant contribution to the sample autocorrelation of FTSE 100 (or, alternatively, to the scalogram). In Z_t, the "standardised" periods of high volatility contribute less to the sample autocorrelation than in the original FTSE 100 series, which would explain why the sample autocorrelation of Z_t exhibits a different dependence structure: it only indicates slight positive dependence at lag 1, but no significant negative dependence at lags 2 or 3.

The above discussion clearly indicates the need for a local analysis of the FTSE 100 data. By looking at the full evolutionary Haar spectrum of FTSE 100, we are able to find out where and how the local autocovariance structure changes over time.

4.5.2 Full evolutionary Haar spectrum analysis

Figure 4.7 shows the estimated evolutionary Haar spectrum of $F'_{t,T}$ = the 2048 last observations of the FTSE 100 index (plotted in Figure 4.1), smoothed using our generic algorithm with spline smoothing. It seems that scale $j = -2$ dominates from time $z_0 \sim 0.6$ onwards (this corresponds, roughly, to time $t = 1200, \ldots, 2048$). In particular, there is a huge bump centred at $z_1 \sim 0.67$: it is clearly the most visible feature in the "spectrum landscape" of FTSE 100. Judging by the magnitude of the bump, it seems likely that even though scale $j = -2$ dominates over part of the time horizon only, "global" tools (such as the scalogram or the sample autocovariance computed for the whole sample) will also be affected, which will give the false impression that scale $j = -2$ dominates all the way through. Indeed, if we compute the acf of $F'_{1,T}, F'_{2,T}, \ldots, F'_{1200,T}$, it turns out that the effect of the sample acf resembling the Haar autocorrelation function at scale $j = -2$, is not present now. The acf of the first 1200 observations of $F'_{t,T}$ is plotted is the left-hand plot of Figure 4.6. Right-hand plot of Figure 4.6 shows the acf of the remaining part of $F'_{t,T}$, where scale $j = -2$ seems to dominate. This is reflected in the shape of the sample acf at lags $1, 2, 3$.

The LSW model with the Haar basis seems to be ideally suited for modelling the FTSE 100 series on the interval $z \in (0.6, 1)$, as it provides a sparse representation of the local covariance in that region: most of the "energy" of the series is concentrated at scales $j = -1$ and -2.

The above demonstrates how important it is to analyse the log-return data *locally*, rather that using global tools. There is no economic reason why log-return series should stay stationary over long periods, and the above wavelet-based analysis shows that, indeed, they do not.

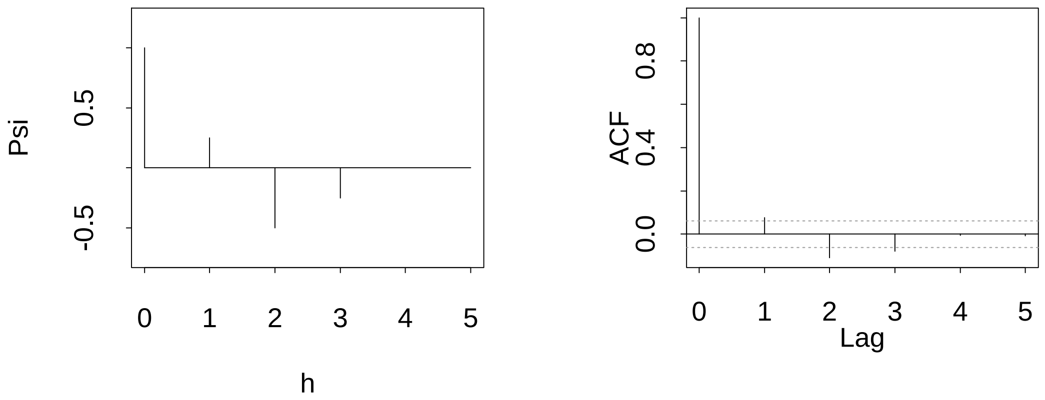

Figure 4.5: Left-hand plot: $\Psi_{-2}(h)$ for Haar wavelets for $h = 0, 1, \ldots, 5$. Right-hand plot: autocorrelation function for $F_{t,T}$ at lags $0, 1, \ldots, 5$.

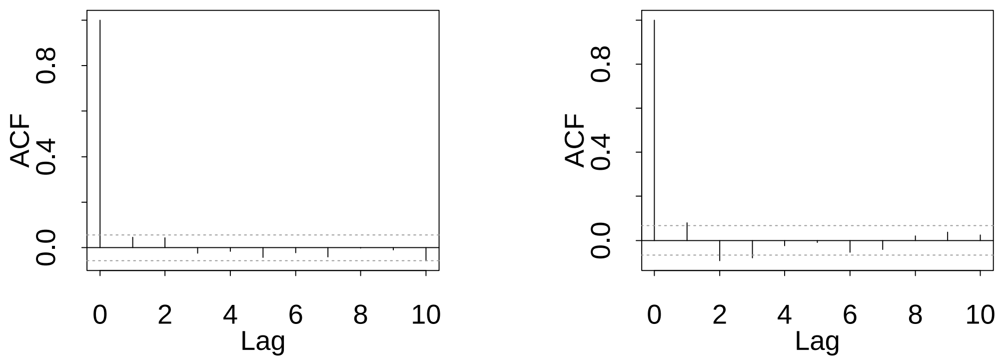

Figure 4.6: Left-hand plot: sample autocorrelation of $F'_{1,T}, \ldots, F'_{1200,T}$. Right-hand plot: sample autocorrelation of $F'_{1201,T}, \ldots, F'_{2048,T}$.

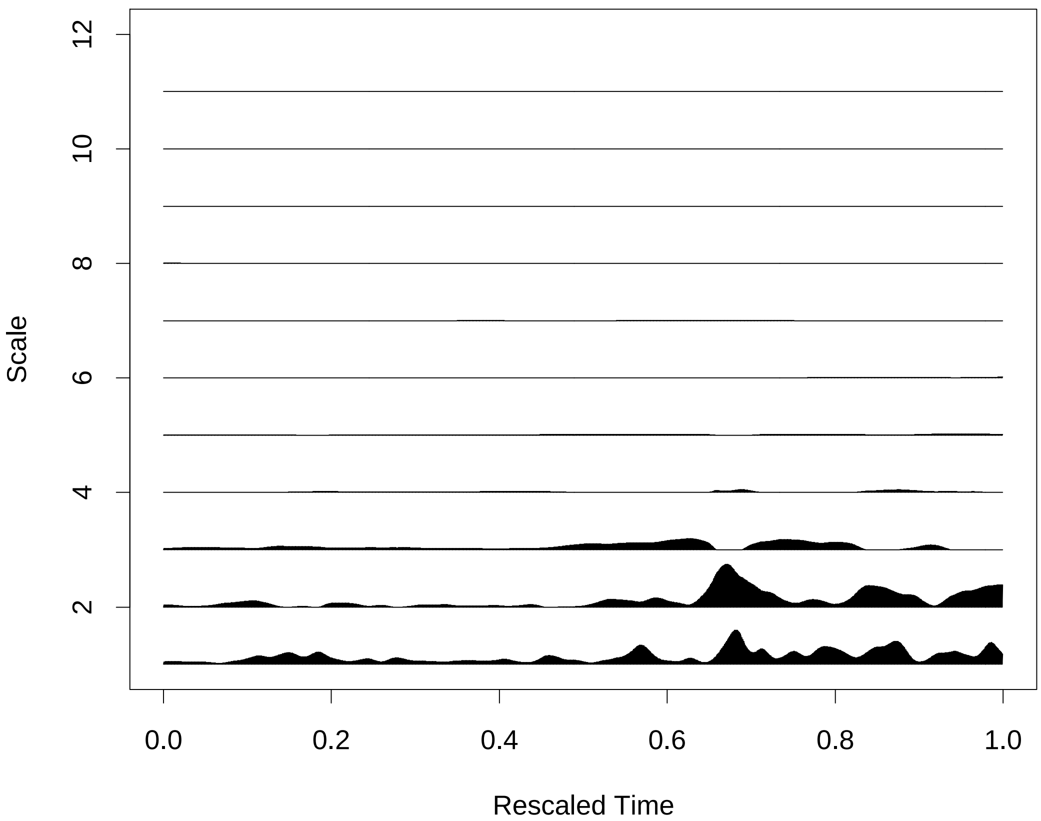

Figure 4.7: Estimated evolutionary Haar spectrum of $T = 2048$ last observations of FTSE 100 of Figure 4.1. Smoothing uses splines. X-axis is the rescaled time $z = t/T$, and Y-axis is negative scale $-j = 1, 2, \ldots, 11$.

4.6 Forecasting

A comparison of forecasting methods for daily Sterling exchange rates is provided by Brooks (1997), who concludes that forecasts based on GARCH modelling are the most reliable. Leung *et al.* (2000) find that probabilistic neural networks (Wasserman (1993)) outperform other methods when applied to stock index returns. However, the input variables in their model include, apart from the past data, a variety of other macroeconomic factors. In this section, we only consider forecasts based on past values of the series, and compare our methodology to forecasting based on GARCH modelling (for an overview of the latter methodology, see e.g. Bera & Higgins (1993)).

The algorithm which we apply here is the adaptive forecasting procedure detailed in Section 3.5 of this thesis. As the theory underlying the algorithm was developed for LSW_2, and not for LSW_3 processes, it would be natural to ask at this point whether a similar theory can also be developed for the latter model. We anticipate that it is indeed the case: we conjecture that using exactly the same techniques, it is possible to obtain results analogous to those of Chapter 3; however, it is likely that the speed of convergence of the relevant quantities will be different. We leave this interesting theoretical problem as a possibility for future research.

We demonstrate the usefulness of the wavelet approach by comparing our forecasting methodology to forecasting based on AR+GARCH modelling, on a fragment of the Dow Jones IA series (denoted by $D_{t,T}$ in Section 4.5 and plotted in Figure 4.2). However, this brief simulation study does not aim to show that our approach is superior to AR+GARCH. Instead, we attempt to demonstrate a few interesting features of LSW_3 forecasting.

Suppose that we have already observed 1105 values of the series, and want to perform one-step prediction of the series along the segment $D_{1106,T}, \dots, D_{1205,T}$. In order to do so, we employ the algorithm of Section 3.5 with Haar wavelets. We make an initial guess at the values of p and g: we set $(p, g) = (1, 30)$. Further, we set the criterion function to

$$d_1\left(\{X_{0,T},\dots,X_{k,T}\},p,g\right)=\left|X_{k,T}-\hat{X}_{k,T}(p,g)\right| \tag{4.28}$$

and we allow one parameter update at each time point.

Also, we limit the parameter space for p to the set $\{1,2\}$, having empirically found that the forecasting algorithm performs best on the given stretch of the series when the upper limit for p is set to 2. This roughly corresponds to "switching" between TMWN and time-varying AR(1) at each time point, depending which model produces *locally* more accurate forecasts.

We compare our method to forecasts obtained by modelling $D_{t,T}$ as

- AR(1) + GARCH(1,1) — since AR(1) roughly corresponds to the upper limit for p being equal to 2,
- AR(16) + GARCH(1,1) — since the AIC criterion indicates that the order of $D_{t,T}$ along the segment $t = 1105,\dots,1204$ is equal to 16.

The parameters $(1,1)$ of the GARCH part were selected *ad hoc*; however, they have no influence on the point forecasts. The models were fitted using the `garch` routine from the S-Plus `garch` module.

The results of the experiment are presented in Figure 4.8. The top left plot shows the actual series $D_{1106,T},\dots,D_{1205,T}$ (dotted line), the corresponding one-step-ahead forecasts (thick solid line), and 95% prediction intervals (assuming Gaussianity; dashed lines), for the AR(1) + GARCH(1,1) model. The top right plot shows the same for the AR(16) + GARCH(1,1) model, and the bottom left plot — the same for the LSW_3 model. The bottom right plot in Figure 4.8 shows the actual series scaled by the factor of 2000 (dotted line), as well as the corresponding values h of the bandwidth used to forecast the series. The bandwidth was allowed to change by ± 1 or remain the same. The fact that it increases steadily beginning from $t = 1160$ may suggest that the time-varying second order structure of $D_{t,T}$ evolves more slowly in that region.

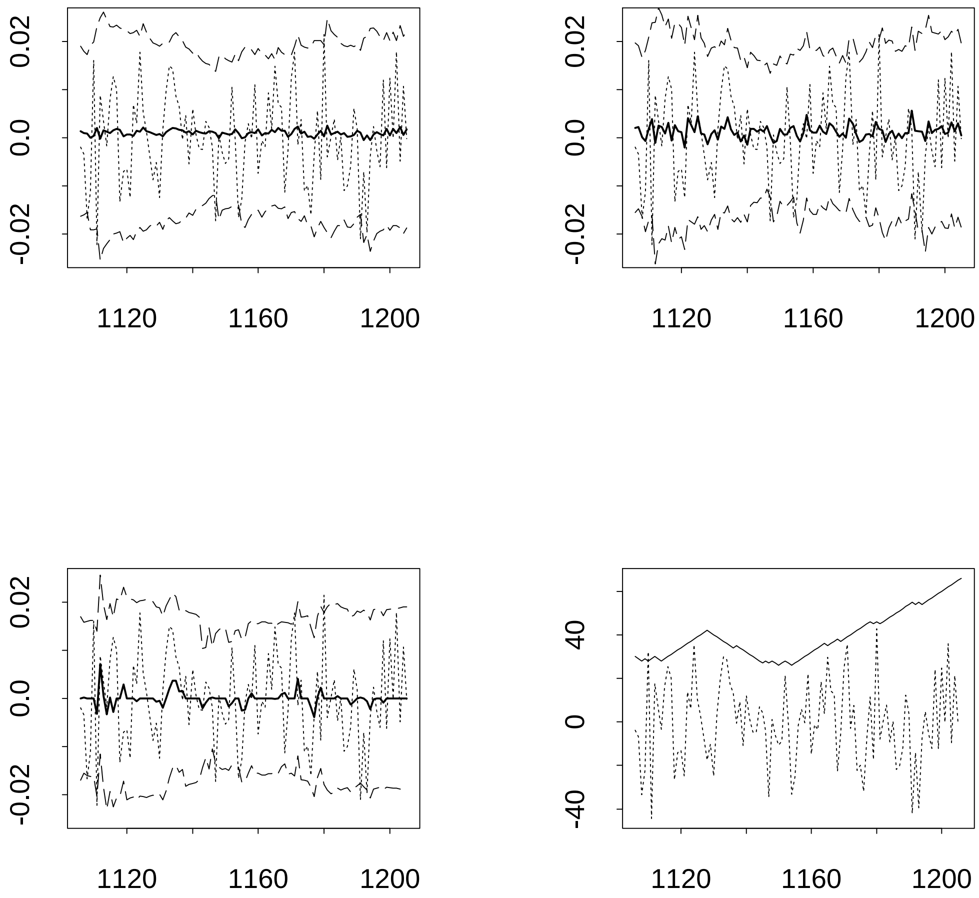

Figure 4.8: Top left, top right and bottom left: the actual series (dotted line), one-step forecasts (solid line) and 95% prediction intervals (dashed lines) for AR(1) + GARCH(1,1), AR(16)+GARCH(1,1) and LSW_3, respectively. Bottom right: actual series $\times 2000$ and the evolution of the bandwidth g.

	AR(1)+GARCH(1,1)	AR(16)+GARCH(1,1)	LSW_3
Mean SPE	878	857	839
Median SPE	404	375	298

Table 4.2: Mean Squared Prediction Error and Median Squared Prediction Error ($\times 10^7$ and rounded) in forecasting $D_{1106,T}, \ldots, D_{1205,T}$ one step ahead, for the three methods tested in Section 4.6.

In the LSW_3 forecasting, the stretches where $p = 1$ wins over $p = 2$ are indicated by one-step forecasts equal to zero (like in TMWN forecasting). Non-zero forecasts indicate that $p = 2$ is used to perform prediction. The LSW_3 model does an impressive job in picking up the spike at $t = 1112$, and also at capturing the local structure around $t = 1135$. The Mean Squared Prediction Errors and the Median Squared Prediction Errors for the three methods are given in Table 4.2: the LSW_3 method outperforms the other two.

For the LSW_3 method, 92% of observations fall within the corresponding one-step 95% prediction intervals, whereas the analogous ratios for the AR(1) + GARCH(1,1) and AR(16) + GARCH(1,1) methods are 94% and 93%, respectively. Our slightly worse performance is due to the fact that the d_1 criterion only minimises the distance between the predicted value and the actual one, and does not take into account the prediction intervals. A modification of the comparison criterion would almost certainly lead to an improvement over the (already good) ratio of 92%.

However, it must be mentioned that the prediction intervals in the LSW_3 model are narrower than the minimum of those in the AR(1) + GARCH(1,1) model and those in the AR(16) + GARCH(1,1) model in 71% of the cases.

We leave the important problem of forecasting volatility in our wavelet-based framework as one of the many possible avenues for future investigation.

4.7 Conclusion

In this chapter, we have provided theoretical and empirical evidence that stock index returns can be successfully modelled and forecast in a time series model which combines wavelets and the concept of rescaled time. Starting from a motivating example of the FTSE 100 series being modelled as a Time-Modulated White Noise (TMWN), we have slightly altered the definition of an LSW process (Nason *et al.* (2000)) so that the altered setup, called LSW_3, includes TMWN as a special case.

We have provided theoretical evidence that the (linear and non-stationary) LSW_3 model can capture the most commonly observed stylised facts. In particular, we have argued that the heavy tails of the marginal distribution, negligible sample autocorrelations, and non-negligible sample autocorrelations of the squares, are all effects which can possibly be caused by applying stationary, global tools (such as the sample autocorrelation) to the analysis of non-stationary data.

Furthermore, we have proposed a new general algorithm for estimating time-varying second-order quantities in the LSW_3 model. We have shown that two particular implementations of our algorithm, specifically designed for financial log-returns, outperform the default algorithm proposed by Nason *et al.* (2000) for general non-stationary time series.

Also, we have provided two interesting examples of exploratory data analysis using the LSW_3 toolbox. By using the (global) scalogram and the (local) evolutionary Haar spectrum, we have found that the daily FTSE 100 index displays a significant local departure from the TMWN setting. Also, by examining the Haar spectrum, and the shape of the autocovariance function of FTSE 100 over a certain region, we have discovered that the Haar wavelet basis is ideally suited for the sparse modelling of FTSE 100 on that interval. The example has powerfully demonstrated that the financial log-return data need to analysed using local tools as all of their second order characteristics, and not only variance, can vary over time.

Finally, we have provided evidence that financial log-returns can be successfully forecast in the LSW_3 framework using the adaptive forecasting algorithm of Section 3.5. We have compared the forecasts obtained by the adaptive algorithm to those obtained using GARCH modelling. Again, we have found that the adaptive method has the potential to accurately forecast some important local features of non-stationary log-return data. In the example analysed (a fragment of the Dow Jones IA index), the LSW_3-based technique has outperformed two GARCH-based methods.

The S-Plus routines written for and used in this chapter, the data sets analysed in it, as well as the contrived standard deviation function of Figure 4.3 are included on the associated CD.

Chapter 5

Denoising the wavelet periodogram using the Haar-Fisz variance stabilising transform

Our aim in this chapter is twofold. Firstly, we introduce a multiscale variance stabilising transform for the wavelet periodogram of a Gaussian LSW process. We call the procedure the "Haar-Fisz" transform, as it consists of three basic steps: taking the Haar transform of the periodogram sequence, dividing the arising detail coefficients by the corresponding smooth coefficients (an instance of the so-called Fisz transform), and finally taking the inverse Haar transform. The resulting vector is closer to Gaussianity than the result of the classical log transform; also, its variance is well stabilised. This is confirmed not only by empirical results but also by theory.

Secondly, we investigate the performance of a denoising method for the wavelet periodogram which consists in taking the Haar-Fisz transform, denoising the transformed vector using a method suitable for Gaussian noise, and then taking the inverse Haar-Fisz transform. Simulations demonstrate excellent performance.

5.1 Motivation: the Fisz transform

The initial motivation for this research was the following result proved in the paper by Fisz (1955) (we stick to the original notation from the paper). Suppose that

$\xi(\lambda)$ is a nonnegative random variable from a family of distributions parametrised by a positive parameter λ, and $m(\lambda) = \mathbb{E}(\xi(\lambda))$, $\sigma^2(\lambda) = \mathrm{Var}(\xi(\lambda))$. We say that $\xi(\lambda)$ is *asymptotically normal* $N(u(\lambda), v(\lambda))$ if there exist functions $u, v > 0$ s.t. for all $x \in \mathbb{R}$,

$$\lim_{\lambda\to\infty} P\left(\frac{\xi(\lambda) - u(\lambda)}{v(\lambda)} < x\right) = \Phi(x), \tag{5.1}$$

where $\Phi(x)$ is the cdf of the standard normal. Let $\xi_i(\lambda_i)$, $i = 1, 2$ be two independent variables and let $m_i = \mathbb{E}(\xi_i)$, $\sigma_i^2 = \mathrm{Var}(\xi_i)$. The following theorem holds.

Theorem 5.1.1 *If*

- $\xi(\lambda)/m(\lambda)$ *converges in probability to one as* $\lambda \to \infty$*;*
- $\xi(\lambda)$ *is asymptotically normal* $N(m(\lambda), \sigma(\lambda))$*;*
- $\lim_{(\lambda_1,\lambda_2)\to(\infty,\infty)} m_1/m_2 = 1$,

then the variable

$$\zeta(\lambda_1, \lambda_2) = \frac{\xi_2(\lambda_2) - \xi_1(\lambda_1)}{(\xi_2(\lambda_2) + \xi_1(\lambda_1))^p}, \tag{5.2}$$

where p *is an arbitrary positive number, is asymptotically normal*

$$N\left(\frac{m_2 - m_1}{(m_2 + m_1)^p}, \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{(m_2 + m_1)^p}\right).$$

Note the specific form of the ratio in (5.2): it can be viewed as the ratio of a Haar detail coefficient and the pth power of the corresponding smooth coefficient. We shall exploit this property later.

As an example, consider $\xi(n) = a(X_1^2 + \ldots + X_n^2)$, where X_i are i.i.d. $N(0, 1)$ and $a > 0$. We have $\xi(n) \sim a\chi_n^2 = a\,\mathrm{Gamma}(\frac{1}{2}, \frac{n}{2})$ and $m(n) = an$, $\sigma^2(n) = 2a^2 n$. The variable $\xi(n)$ satisfies the first two assumptions of Theorem 5.1.1 by the Law of Large Numbers and the Central Limit Theorem, respectively. Assume that $\xi_1(n_1)$, $\xi_2(n_2)$ are independent and $n_1/n_2 \to 1$. Then, $\zeta(n_1, n_2)$ is asymptotically normal

$$N\left(\frac{n_2 - n_1}{(n_2 + n_1)^p} a^{1-p}, \frac{\sqrt{2}}{(n_2 + n_1)^{p-1/2}} a^{1-p}\right).$$

Note that setting $p = 1$ makes the variance of $\zeta(n_1, n_2)$ independent of a. Indeed, this *variance stabilisation* is the key property of the Haar-Fisz transform of Section 5.3. We now formally define the *Fisz transform*, an essential component of the Haar-Fisz transform.

Definition 5.1.1 *Let X, Y be two nonnegative random variables. The* Fisz transform of X and Y with exponent p *is defined as*

$$\zeta_p(X, Y) = \frac{X - Y}{(X + Y)^p}, \tag{5.3}$$

with the convention that $0/0 = 0$.

5.2 Properties of the wavelet periodogram in the Gaussian LSW model

As mentioned above, the Fisz transform can be viewed as the division of a Haar detail coefficient by the pth power of the corresponding smooth coefficient (up to a multiplicative constant). In actuality, we are interested in applying this operation to Haar coefficients of wavelet periodogram sequences in the Gaussian LSW model. Therefore, let us now recall some properties of this statistic.

For Gaussian LSW processes, the wavelet periodogram $I_{t,T}^{(j)}$ at a fixed scale j is a sequence of scaled χ_1^2 variables. Also, we know from Section 2.2.4 that $I_{t,T}^{(j)}$ is an asymptotically unbiased, but inconsistent estimator of $\beta_j(t/T)$, where

$$\beta_j(z) := \sum_{i=-\infty}^{-1} S_i(z) A_{i,j}. \tag{5.4}$$

Furthermore, the following proposition shows that the wavelet periodogram at each scale j is typically a correlated sequence.

Proposition 5.2.1 *Let $X_{t,T}$ be a Gaussian LSW process satisfying $S_j(z) \leq D2^j$. We have*

$$\operatorname{cov}\left(I_{t,T}^{(j)}, I_{t+s,T}^{(j)}\right) = 2\left(\sum_{\tau=-\infty}^{\infty} c\left(\frac{t}{T}, \tau\right) \Psi_j(\tau + s)\right)^2 + O(2^{-j}/T). \tag{5.5}$$

The proof uses exactly the same technique as the proof of (2.38).

Our ultimate objective is to denoise the periodogram sequences at scales $j = -1, \ldots, -J(T)$, i.e. to provide estimates of the functions $\beta_{-1}(z), \ldots, \beta_{-J(T)}(z)$. Being able to estimate $\{\beta_j(z)\}_j$ is useful in two ways:

1. Estimates of $\{\beta_j(z)\}_j$ can be used to obtain estimates of $\{S_j(z)\}_j$ (by (5.4) and by the invertibility of A, see Nason *et al.* (2000) for details);

2. The estimate of $S_j(z)$ can in turn be used to obtain an estimate of the local autocovariance $c(z, \tau)$ (using the representation (2.32)).

In short, estimating $\{\beta_j(z)\}_{j=-1}^{-J(T)}$ allows us to make inference about the time-varying second-order structure of $X_{t,T}$.

The top plot in Figure 5.1 shows an example of the wavelet spectrum $S_j(z)$ where only $S_{-1}(z)$ and $S_{-3}(z)$ are non-zero. The middle plot shows a sample path of length 1024 simulated from it, using Haar wavelets and Gaussian innovations. The bottom plot shows the Haar periodogram of the simulated series at scale -1.

Denoising the wavelet periodogram is by no means an easy task, due to

- the fact that the variance of the noise depends on the level of the signal (see formulas (2.37) and (2.38)),

- an extremely low signal-to-noise ratio: again by (2.37) and (2.38) we obtain, neglecting the remainders, $\mathbb{E}I_{t,T}^{(j)} / \left\{\mathrm{Var}\left(I_{t,T}^{(j)}\right)\right\}^{1/2} = 2^{-1/2}$,

- the presence of correlation in the noise (see formula (5.5)).

Most existing denoising techniques have been designed to handle stationary Gaussian noise and therefore it would be desirable to be able to transform the wavelet periodogram into a signal contaminated with such noise before the denoising is performed. A well-known technique for stabilising the variance of scaled χ_n^2 variables is the log-transform, see e.g. Priestley (1981); however, the resulting variable is still far from Gaussian if, like here, $n = 1$. Nason *et al.* (2000) propose a wavelet-based technique for denoising the wavelet periodogram without any pre-processing.

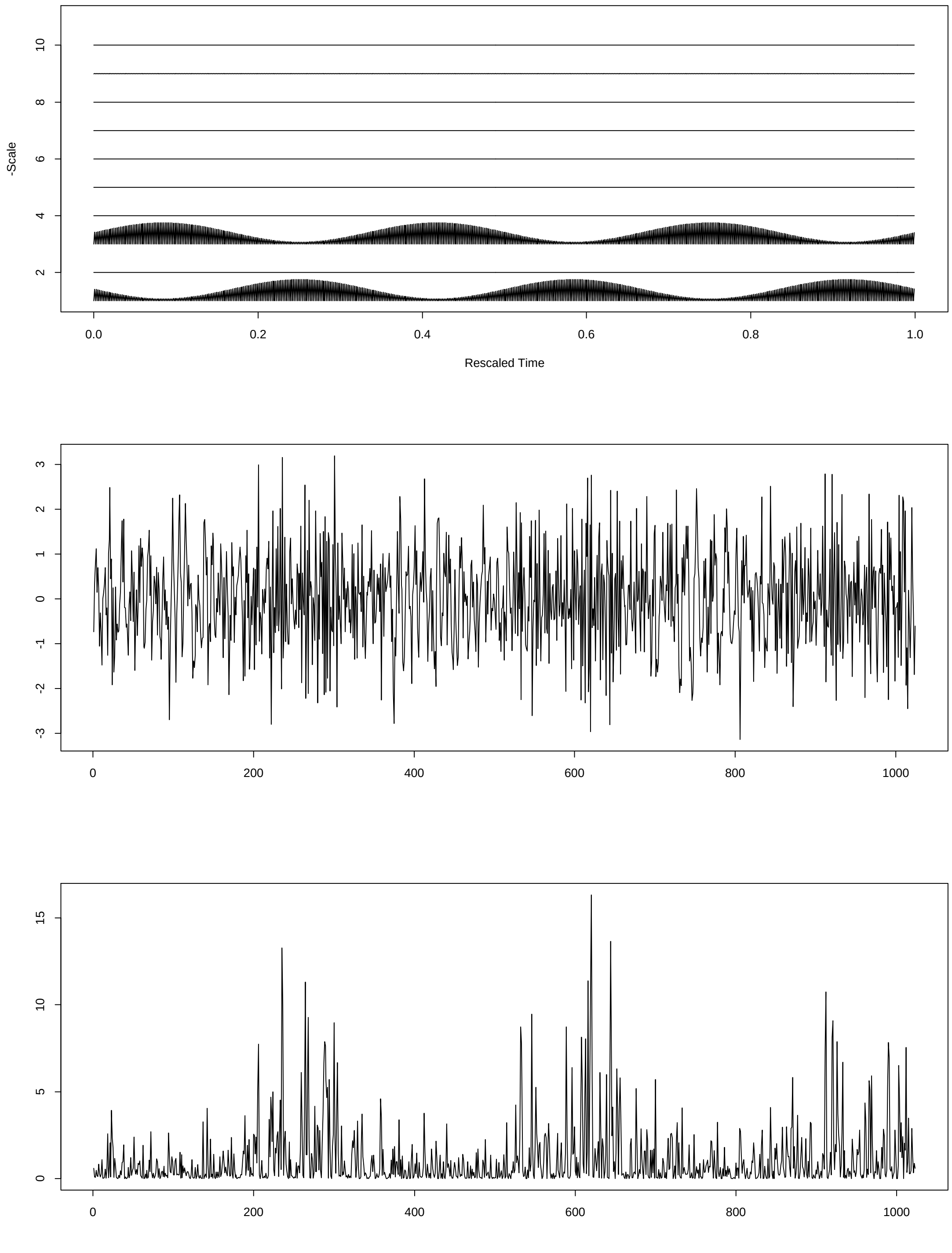

Figure 5.1: Top plot: example of a wavelet spectrum where only $S_{-1}(z)$ and $S_{-3}(z)$ are non-zero. Middle plot: sample path of length 1024 simulated from this spectrum using Haar wavelets and Gaussian innovations. Bottom plot: the Haar periodogram of the simulated realisation at scale $j = -1$.

In the next section, we introduce the Haar-Fisz transform: a multiscale Gaussianising and variance stabilising transformation for the wavelet periodogram which turns out to be a viable alternative for the log transform.

5.3 The Haar-Fisz transform

5.3.1 Algorithm for the Haar-Fisz transform

In this section, we provide details of the Haar-Fisz transform, which stabilises the variance of the wavelet periodogram and brings its distribution closer to normality. The input to the algorithm is:

- A single row of the wavelet periodogram $I_{t,T}^{(j)}$ at a fixed scale j: here, we assume that T is an integer power of two. To simplify the notation in this section, we drop the superscript j and the subscript T and consider the sequence $I_t := I_{t,T}^{(j)}$, or, in vector notation, $\mathbf{I} = (I_{0,T}^{(j)}, \dots, I_{T-1,T}^{(j)})'$.
- A fixed integer $M \in \{1, 2, \dots, \log_2(T)\}$; its meaning will become clear later.

The output from the algorithm is:

- The mean of $\mathbf{I}$, denoted by $\bar{\mathbf{I}}$.
- A vector $\mathbf{U}^M$ of length 2^M.

The vector $\mathbf{U}^M$ is constructed as follows:

1. Let $\mathbf{s}^M$ be the vector of local averages of $\mathbf{I}$:

$$s_n^M = \frac{2^M}{T} \sum_{t=nT2^{-M}}^{(n+1)T2^{-M}-1} I_t \quad \text{for} \quad n = 0, 1, \dots, 2^M - 1. \tag{5.6}$$

2. For each $m = M-1, M-2, \dots, 0$, recursively form vectors $\mathbf{s}^m$ and $\mathbf{f}^m$:

$$s_n^m = \frac{1}{2}(s_{2n}^{m+1} + s_{2n+1}^{m+1}) \tag{5.7}$$

$$f_n^m = \frac{s_{2n}^{m+1} - s_{2n+1}^{m+1}}{2 s_n^m}, \tag{5.8}$$

   for $n = 0, 1, \dots, 2^m - 1$, with the convention $0/0 = 0$.

3. For each $m = 0, 1, \dots, M-1$, recursively modify the vectors $\mathbf{s}^{m+1}$:

$$s_{2n}^{m+1} = s_n^m + f_n^m \tag{5.9}$$

$$s_{2n+1}^{m+1} = s_n^m - f_n^m, \tag{5.10}$$

   for $n = 0, 1, \dots, 2^m - 1$.

4. Set $\mathbf{U}^M := \mathbf{s}^M - \bar{\mathbf{I}}$.

We denote $\mathcal{F}^M \mathbf{I} := \mathbf{U}^M$. The nonlinear operator $\mathcal{F}^M$ is called the *Haar-Fisz transform of* $\mathbf{I}$ *at the resolution level* M.

If $M = \log_2(T)$, then the length of $\mathcal{F}^M \mathbf{I}$ is T and the algorithm is invertible, i.e. $\mathbf{I}$ can be reconstructed from $\mathcal{F}^M \mathbf{I}$ and $\bar{\mathbf{I}}$ by reversing the steps 4.–1. Therefore, the case $M = \log_2(T)$ is the one we are the most interested in in practice. However, the exact asymptotic Gaussianising properties of the Haar-Fisz transform only hold for M fixed (i.e. independent of T), and this case is investigated theoretically in Section 5.5.1. Section 5.5.2 provides some heuristics as to the behaviour of $\mathcal{F}^M \mathbf{I}$ when $M = \log_2(T)$: we still conclude that the distribution of $\mathcal{F}^{\log_2(T)} \mathbf{I}$ is close to Gaussian with a constant variance. To simplify notation, we denote $\mathcal{F} := \mathcal{F}^{\log_2(T)}$.

Note that the steps 2.–4. of the algorithm are similar to the forward and inverse Discrete Haar Transform except the division by s_n^m in formula (5.8). Also, observe that (5.8) can be written as

$$f_n^m = \zeta_1(s_{2n}^{m+1}, s_{2n+1}^{m+1}). \tag{5.11}$$

That is, f_n^m is the result of the Fisz transform with exponent 1 of two neighbouring smooth coefficients s_{2n}^{m+1} and s_{2n+1}^{m+1}.

Finally, note that the Haar-Fisz transform, being a *computationally* straightforward modification of the Haar transform, is also of computational order O(T).

5.3.2 Examples

As an example, consider $T = 8$. For $M = 2$, we have:

$$\begin{aligned}
U_0^2 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} \\
U_1^2 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} \\
U_2^2 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t} \\
U_3^2 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t}.
\end{aligned}$$

Similarly, for $M = 3$, we have:

$$\begin{aligned}
U_0^3 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} + \frac{I_0 - I_1}{I_0 + I_1} \\
U_1^3 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} - \frac{I_0 - I_1}{I_0 + I_1} \\
U_2^3 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} + \frac{I_2 - I_3}{I_2 + I_3} \\
U_3^3 &= \frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_0 + I_1 - I_2 - I_3}{\sum_{i=0}^{3} I_t} - \frac{I_2 - I_3}{I_2 + I_3} \\
U_4^3 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t} + \frac{I_4 - I_5}{I_4 + I_5} \\
U_5^3 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} + \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t} - \frac{I_4 - I_5}{I_4 + I_5} \\
U_6^3 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t} + \frac{I_6 - I_7}{I_6 + I_7} \\
U_7^3 &= -\frac{\sum_{t=0}^{3} I_t - \sum_{t=4}^{7} I_t}{\sum_{i=0}^{7} I_t} - \frac{I_4 + I_5 - I_6 - I_7}{\sum_{i=4}^{7} I_t} - \frac{I_6 - I_7}{I_6 + I_7}
\end{aligned}$$

Figure 5.2 compares the log transform (left plot) and the Haar-Fisz transform (right plot) of the wavelet periodogram from the bottom plot of Figure 5.1. Here, $T = 1024$, and the full transform is performed, i.e. $M = 10$. The Haar-Fisz-transformed wavelet periodogram appears to be much closer to normality.

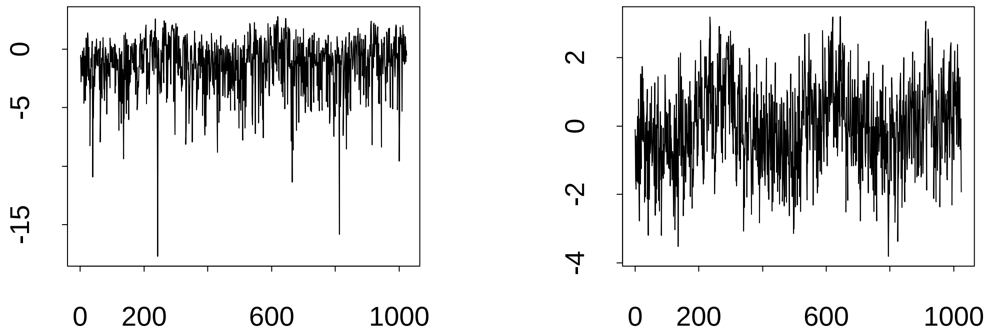

Figure 5.2: The log transform (left plot) and the Haar-Fisz transform with $M = 10$ (right plot) of the wavelet periodogram from the bottom plot of Figure 5.1.

5.4 A Functional CLT for the centred wavelet periodogram

In this section, we are concerned with a Functional Central Limit Theorem (FCLT) for the centred wavelet periodogram $Z_{t,T}^{(j)} := I_{t,T}^{(j)} - \mathbb{E}I_{t,T}^{(j)}$ (see Davidson (1994) for more on the stochastic limit theory we use here). Our FCLT demonstrates that the normalised cumulative sum of the centred wavelet periodogram converges in distribution to a transformed Brownian motion. The theory in this section enables us to demonstrate the Gaussianising, variance stabilising and decorrelating properties of the Haar-Fisz transform established in Section 5.5. Before we state the theorem, we introduce some essential notation.

Definition 5.4.1 (transformed Brownian motion) *Let η be an increasing homeomorphism on $[0,1]$ with $\eta(0) = 0$ and $\eta(1) = 1$. A transformed Brownian motion B_η is defined as*

$$B_\eta(z) \stackrel{D}{=} B(\eta(z)), \quad z \in [0,1],$$

where B is the standard Brownian motion.

Definition 5.4.2 (cross-scale autocorrelation wavelets) *Let ψ be a fixed wavelet system. Vectors $\Psi_{i,j}$, for $i, j \in \{-1, -2, \dots\}$, defined by*

$$\Psi_{i,j}(\tau) = \sum_{s=-\infty}^{\infty} \psi_{i,s+\tau} \psi_{j,s} \tag{5.12}$$

are called the cross-scale autocorrelation wavelets.

Denote

$$\begin{aligned} \bar{S}_j &= \max_z S_j(z) \\ \delta_k &= \max_{j=-1,\dots,-k} \frac{C_j}{\bar{S}_j}, \end{aligned}$$

with the convention $0/0 = 0$. Denote further

$$\begin{aligned} A^{\tau}_{i,j} &= \sum_n \Psi_{i,j}(n) \Psi_{i,j}(n+\tau) \\ \beta^{\tau}_j(z) &= \sum_i S_i(z) A^{\tau}_{i,j} \end{aligned} \tag{5.13}$$

We now state the Functional Central Limit Theorem for the centred wavelet periodogram.

Theorem 5.4.1 *Let $X_{t,T}$ be a Gaussian LSW process, and let $Z^{(j)}_{t,T}$ be its centred wavelet periodogram at scale j. Define*

$$\begin{aligned} b^2_T &= \mathbb{E} \left(\sum_{t=0}^{T-1} Z^{(j)}_{t,T} \right)^2 \\ R_T(z) &= \frac{\sum_{t=0}^{\lfloor zT \rfloor - 1} Z^{(j)}_{t,T}}{b_T} \quad \text{for} \quad z \in [0,1]. \end{aligned}$$

If

$$\exists\, \varepsilon > 0 \quad \left(\sum_{i<0} \sum_{l \geq m+1} \Psi^2_{i,j}(l) \bar{S}_i \right)^{1/2} = O(m^{-1/2-\varepsilon}), \tag{5.14}$$

$$\delta_{J(T)}/T \in l_{\infty}, \tag{5.15}$$

$$\sup_z \sum_n |c(z,n)| < \infty, \tag{5.16}$$

$$\exists\, D \quad \bar{S}_j 2^{-j} \leq D \quad \forall\, j, \tag{5.17}$$

then $R_T \xrightarrow{D} B_{\eta}$, where

$$\eta(z) = \frac{\int_0^z \sum_{\tau=-\infty}^{\infty} \left(\sum_i S_i(u) A^{\tau}_{i,j} \right)^2 du}{\int_0^1 \sum_{\tau=-\infty}^{\infty} \left(\sum_i S_i(u) A^{\tau}_{i,j} \right)^2 du}.$$

The proof of Theorem 5.4.1 appears later in this section. As is clear from the proof, the left-hand expression in condition (5.14) is a measure of dependence in the sequence $Z_{t,T}^{(j)}$ at lag m. Condition (5.15) places an additional restriction on the finite-sample wavelet spectrum $\{\omega_{j,k;T}^2\}_{j,k}$ in relation to the asymptotic spectrum $\{S_j(z)\}_j$. Condition (5.16) is a short-memory assumption for $X_{t,T}$, and condition (5.17) requires that the wavelet spectrum should decay at a certain speed as $j \to -\infty$.

We now give an example of a periodogram sequence which satisfies the technical condition (5.14). Let $X_{t,T}$ be a Gaussian LSW process constructed with Haar wavelets and such that $S_i(z) = S_i = 2^i$. Asymptotically, $X_{t,T}$ is a white noise process (see formula 3.70). Consider the Haar periodogram of $X_{t,T}$ at scale $j = -1$. Using the explicit form of discrete Haar vectors and (5.12), simple algebra yields

$$\sum_{i<0} 2^i \sum_{l \geq m+1} \Psi_{i,-1}^2(l) = O(m^{-2}),$$

and (5.14) is satisfied.

In particular, Theorem 5.4.1 implies that $\mathbb{E}(R_T(z)^2) \to \eta(z)$ as $T \to \infty$, and that the increments of $R_T(z)$ are asymptotically independent. Theorem 5.4.1 is fundamental for the theoretical results of the next section.

To be able to prove Theorem 5.4.1, we first need to recall the definition of L_2-Near Epoch Dependence (L_2-NED), and then prove two technical lemmas.

Definition 5.4.3 *For a stochastic array $\{\{V_{t,T}\}_{t=-\infty}^{\infty}\}_{T=1}^{\infty}$, possibly vector-valued, on a probability space $(\Omega, \mathcal{G}, P)$, let $\mathcal{G}_{t-m,T}^{t+m} = \sigma(V_{t-m,T}, \ldots, V_{t-m,T})$. If an integrable array $\{\{X_{t,T}\}_{t=-\infty}^{\infty}\}_{T=1}^{\infty}$ satisfies*

$$\|X_{t,T} - \mathbb{E}(X_{t,T}|\mathcal{G}_{t-m,T}^{t+m})\|_2 \leq h_{t,T}\nu_m,$$

where $\nu_m \to 0$, and $\{h_{t,T}\}$ is an array of positive constants, it is said to be L_2-NED on $\{V_{t,T}\}$ with constants $\{h_{t,T}\}$. Further, if $\exists\, \varepsilon > 0$ s.t. $\nu_m = O(m^{\lambda-\varepsilon})$, then $\{\{X_{t,T}\}_{t=-\infty}^{\infty}\}_{T=1}^{\infty}$ is said to be L_2-NED of size λ on $\{V_{t,T}\}$.

Lemma 5.4.1 *Define $\underline{\xi}_{t,T} = (\xi_{-1,t}, \ldots, \xi_{-J(T),t})'$. If*

$$\exists\, \varepsilon > 0 \quad \left(\sum_{i<0} \sum_{l \geq m+1} \Psi_{i,j}^2(l) \bar{S}_i \right)^{1/2} = O(m^{-1/2-\varepsilon}),$$

then $Z_{t,T}^{(j)}/b_T$ *is* L_2*-NED of size* $-1/2$ *on* $\{\underline{\xi}_{t,T}\}$*. If in addition* $\delta_{J(T)}/T \in l_\infty$*, then the NED constants can be set to* $1/b_T$*.*

Proof. It suffices to examine the L_2-Near Epoch Dependence for $Z_{t,T}^{(j)}$. Define

$$\mathcal{G}_{t-m,T}^{t+m} = \sigma(\underline{\xi}_{t-m,T}, \dots, \underline{\xi}_{t+m,T}).$$

We have

$$\begin{aligned}
Z_{t,T}^{(j)} - \mathbb{E}(Z_{t,T}^{(j)}|\mathcal{G}_{t-m,T}^{t+m}) &= I_{t,T}^{(j)} - \mathbb{E}(I_{t,T}^{(j)}|\mathcal{G}_{t-m,T}^{t+m}) \\
&= \left| \sum_{i=-1}^{-J(T)} \sum_k \omega_{i,k;T} \Psi_{i,j}(t-k)\xi_{i,k} \right|^2 \\
&- \left| \sum_{i=-1}^{-J(T)} \sum_{k \geq t-m} \omega_{i,k;T} \Psi_{i,j}(t-k)\xi_{i,k} \right|^2 \\
&- \sum_{i=-1}^{-J(T)} \sum_{k<t-m} \omega_{i,k;T}^2 \Psi_{i,j}^2(t-k) \\
&= Y_1^2 - Y_2^2 - K_1^2 = (Y_1 - Y_2)(Y_1 + Y_2) - K_1^2,
\end{aligned}$$

where Y_n^2 are random and K_1 is deterministic. Note that $Y_1 - Y_2$ and $Y_1 + Y_2$ are Gaussian and that $\mathbb{E}(Y_1 - Y_2)^2 = \mathbb{E}((Y_1 - Y_2)(Y_1 + Y_2)) = K_1^2$. Simple algebra yields

$$\mathbb{E}\left((Y_1 - Y_2)(Y_1 + Y_2) - K_1^2\right)^2 = 2K_1^2 \mathbb{E}(Y_1^2 + Y_2^2) \leq 4K_1^2 \mathbb{E}Y_1^2.$$

Noting that

$$\begin{aligned}
K_1^2 &\leq \left(1 + \frac{\delta_{J(T)}}{T}\right) \sum_{i=-1}^{-\infty} \sum_{l \geq m+1} \Psi_{i,j}^2(l)\bar{S}_i \\
\mathbb{E}Y_1^2 &\leq \left(1 + \frac{\delta_{J(T)}}{T}\right) \sum_{i=-1}^{-\infty} \sum_{l} \Psi_{i,j}^2(l)\bar{S}_i,
\end{aligned}$$

and recalling that $\delta_{J(T)}/T \in l_\infty$, the assertion of the Lemma follows. □

Lemma 5.4.2 *If*

$$\sup_{z\in[0,1]} \sum_\tau |c(z,\tau)| < \infty \tag{5.18}$$

$$\exists D \quad \bar{S}_j 2^{-j} \leq D \quad \forall j, \tag{5.19}$$

then

$$\frac{b_T^2}{T} \to 2\int_0^1 \sum_{\tau=-\infty}^{\infty} \left(\sum_i S_i(z) A_{i,j}^{\tau} \right)^2 dz$$

as $T \to \infty$.

Proof. Using Gaussianity, we have

$$\begin{aligned}
b_T^2 &= 2\sum_{t=0}^{T-1} \sum_{\tau=-t}^{T-1-t} \left(\sum_{i=-1}^{-J(T)} \sum_k \omega_{i,k;t}^2 \Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k) \right)^2 \\
&= 2\sum_{t=0}^{T-1} \sum_{\tau=-t}^{T-1-t} \left(\sum_{i=-1}^{-J(T)} \sum_k \left\{ S_i\left(\frac{t}{T}\right) + O\left(\frac{C_i + L_i(t-k)}{T}\right) \right\} \right. \\
&\times \left. \Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k) \right)^2 \\
&= 2\sum_{t=0}^{T-1} \sum_{\tau=-t}^{T-1-t} \left(\sum_i S_i\left(\frac{t}{T}\right) A_{i,j}^{\tau} \right)^2 + \mathrm{Rest}_T,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{Rest}_T &= 2\sum_{t=0}^{T-1} \sum_{\tau=-t}^{T-1-t} \left(\sum_{i=-1}^{-J(T)} \sum_k O\left(\frac{C_i + L_i(t-k)}{T}\right) \Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k) \right. \\
&- \left. \sum_{i=-J(T)-1}^{-\infty} S_i\left(\frac{t}{T}\right) A_{i,j}^{\tau} \right) \\
&\times \left(\sum_{i=-1}^{-J(T)} \sum_k \left\{ 2S_i\left(\frac{t}{T}\right) + O\left(\frac{C_i + L_i(t-k)}{T}\right) \right\} \right. \\
&\times \left. \Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k) + \sum_{i=-J(T)-1}^{-\infty} S_i\left(\frac{t}{T}\right) A_{i,j}^{\tau} \right). \qquad (5.20)
\end{aligned}$$

Let us first show two simple auxiliary results.

1. Summability of constants C_i and L_i. We use the properties of A from Lemma 3.4.1.

$$\begin{aligned}
\sum_i (C_i + L_i(2^{-i}+2^{-j}))A_{i,j} &= \sum_i (C_i + L_i 2^{-i}) 2^j 2^{-j} A_{i,j} \\
&+ \sum_i L_i 2^{-j} 2^i 2^{-i} A_{i,j} \\
&\leq 2^{-j} \sum_i (C_i + L_i 2^{-i}) \sum_k 2^k A_{i,k} \\
&+ 2^{-j} \sum_i L_i 2^{-i} \sum_k 2^k A_{k,j} \\
&= O(2^{-j}). \qquad (5.21)
\end{aligned}$$

2. Summability of covariance of wavelet coefficients.

$$\begin{aligned}
\sum_\tau \left| \sum_i S_i(z) A^\tau_{i,j} \right| &= \sum_\tau \left| \sum_i S_i(z) \sum_n \Psi_i(n)\Psi_j(n+\tau) \right| \\
&= \sum_\tau \left| \sum_n c(z,n)\Psi_j(n+\tau) \right| \\
&\leq \sum_n |c(z,n)| \sum_\tau |\Psi_j(n+\tau)| \\
&\leq K_1 2^{-j} \sum_n |c(z,n)| \\
&= O(2^{-j}), \qquad (5.22)
\end{aligned}$$

by assumption (5.18).

By formula (5.21) and assumption (5.19), we have

$$\begin{aligned}
&\max_{t,\tau} \left| \sum_{i,k} O\left(\frac{C_i + L_i(t-k)}{T}\right) \Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k) - \sum_{i=-J(T)-1}^{-\infty} S_i\left(\frac{t}{T}\right) A^\tau_{i,j} \right| \leq \\
&O(T^{-1}) \max_{t,\tau} \sum_i C_i + ML_i(2^{-i}+2^{-j}) \sum_k |\Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k)| + \\
&+ \max_{t,\tau} \sum_k \sum_{i=-J(T)-1}^{-\infty} \bar{S}_i \Psi_j(k) \leq \\
&O(T^{-1}) \sum_i \left(C_i + ML_i(2^{-i}+2^{-j})\right) A_{i,j} + O(2^{-j}T^{-1}) = O(2^{-j}T^{-1}). \qquad (5.23)
\end{aligned}$$

Using first (5.23), and then (5.22) and (5.21), we bound (5.20) as follows

$$\begin{aligned}
\mathrm{Rest}_T &\leq O(2^{-j}T^{-1})\sum_{t=0}^{T-1}\sum_{\tau=-t}^{T-1-t}\left|\sum_{i,k}\left\{3S_i\left(\frac{t}{T}\right)+O\left(\frac{C_i+L_i(t-k)}{T}\right)\right\}\right. \\
&\times \left.\Psi_{i,j}(t-k)\Psi_{i,j}(t+\tau-k)\right| \\
&\leq O(2^{-j}T^{-1})\sum_{t=0}^{T-1}\sum_{\tau=-t}^{T-1-t}\left|\sum_i S_i\left(\frac{t}{T}\right)A_{i,j}^{\tau}\right| \\
&+ O(2^{-j}T^{-2})\sum_{t=0}^{T-1}\sum_{\tau=-t}^{T-1-t}\sum_i (C_i+L_i(2^{-i}+2^{-j}))A_{i,j} \\
&= O(2^{-2j})+O(2^{-2j}),
\end{aligned}$$

which yields the result. □

Proof of Theorem 5.4.1. We apply Theorem 29.14 from Davidson (1994), with

$$U_{t,T} = Z_{t,T}^{(j)}/b_T \tag{5.24}$$

$$c_{t,T} = 1/b_T \tag{5.25}$$

$$K_T(z) = \lfloor zT \rfloor, \tag{5.26}$$

where the LHS's of (5.24) – (5.26) use the notation from Davidson (1994), and the RHS's of these formulas use the notation from this chapter. We now check conditions (a) — (f) from Davidson (1994).

(a) Clearly, $\mathbb{E}Z_{t,T}^{(j)} = 0$.

(b) For Gaussian LSW processes, we have $\sup_{t,T}\|Z_{t,T}^{(j)}\|_r < \infty$ for $r > 2$.

(c) Satisfied by Lemma 5.4.1 as $\{\underline{\xi}_{t,T}\}$ independent.

(d) Satisfied by Lemma 5.4.2 as

$$\limsup_{T\to\infty}\frac{\sum_{t=\lfloor zT\rfloor}^{\lfloor (z+\omega)T\rfloor-1}\frac{1}{b_T^2}}{\omega} = \limsup_{T\to\infty}\frac{T}{b_T^2}.$$

(e) We clearly have $1/b_T = O(T^{1/2-1}) = O(T^{-1/2})$.

(f) Again by Lemma 5.4.2, we have

$$\mathbb{E}R_T^2(z) = \frac{b^2_{\lfloor zT \rfloor}}{b_T^2} \to \eta(z).$$

This completes the proof. □

5.5 Properties of the Haar-Fisz transform

5.5.1 Properties of the Haar-Fisz transform for M fixed

In this section, we quantify the Gaussianising, variance stabilising and decorrelating properties of the Haar-Fisz transform for M fixed. The following theorem holds:

Theorem 5.5.1 *Let $X_{t,T}$ satisfy the assumptions of Theorem 5.4.1, and let $I_{t,T}^{(j)}$ be the wavelet periodogram of $X_{t,T}$ at scale j. Let the corresponding functions $\beta_j(z)$ and $\sum_\tau(\beta_j^\tau(z))^2$ (see formulas (5.4) and (5.13)) be continuous with bounded one-sided derivatives. Further, let $\beta_j(z)$ be bounded away from zero. For M fixed, $\mathbf{U}^M = \mathcal{F}^M I_{t,T}^{(j)}$ admits the following decomposition:*

$$\mathbf{U}^M = \mathbf{V}^M + \mathbf{Y}^M,$$

where

1. *$\mathbf{V}^M$ has an almost-sure deterministic limit as $T \to \infty$;*
2. *$\sqrt{T}\mathbf{Y}^M \xrightarrow{D} N(0, \Sigma)$ as $T \to \infty$, with*

$$\begin{aligned}\left(2^{M+1}-2\right) \inf_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} - O(M) &\le \Sigma_{n,n} \\ \le \left(2^{M+1}-2\right) \sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} &+ O(M) \qquad (5.27)\end{aligned}$$

and

$$\Sigma_{n_1,n_2} = O(M) \quad \text{for} \quad n_1 \neq n_2. \qquad (5.28)$$

Property 2. above is called the *Gaussianisation property* of the Haar-Fisz transform. Formulas (5.27) and (5.28) define, respectively, the *variance stabilisation* and the *decorrelation properties* of the Haar-Fisz transform. Note the different asymptotic regimes for $\mathbf{V}^M$ and $\mathbf{Y}^M$: the multiplication of $\mathbf{Y}^M$ by $\sqrt{T}$ is needed because $\mathrm{Var}(Y_n^M) = O(2^M/T)$; remember that M is fixed. However, for the invertible case (see the discussion in Section 5.3), we require $M = \log_2(T)$. Even though this case is extremely challenging to investigate theoretically, we cast some light on the behaviour of $\mathcal{F}^{\log_2(T)}$ in Section 5.5.2. We now prove Theorem 5.5.1.

Proof. Denote $Z_t = I_t - \mathbb{E}I_t$ and recall that $\beta_j^\tau(z) = \sum_i S_i(z)A_{i,j}^\tau$. Note that $\beta_j^0(z) = \beta_j(z)$. Consider a single Haar-Fisz summand f_n^m, for $m \in \{0, 1, \dots, M-1\}$ and $n \in \{0, 1, \dots, 2^m - 1\}$. In what follows, $\gamma_{m,n}$ are appropriate integers and $\alpha_{m,n} \in \{0, 1\}$. We have

$$
\begin{aligned}
f_n^m &= (-1)^{\alpha_{m,n}} \frac{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+1)T2^{-(m+1)}-1} I_t - \sum_{t=(\gamma_{m,n}+1)T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} I_t}{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} I_t} \\
&= (-1)^{\alpha_{m,n}} \frac{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+1)T2^{-(m+1)}-1} Z_t - \sum_{t=(\gamma_{m,n}+1)T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} Z_t}{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} Z_t + \sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} \mathbb{E}I_t} \\
&+ (-1)^{\alpha_{m,n}} \frac{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+1)T2^{-(m+1)}-1} \mathbb{E}I_t - \sum_{t=(\gamma_{m,n}+1)T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} \mathbb{E}I_t}{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} Z_t + \sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} \mathbb{E}I_t} \\
&= y_n^m + v_n^m.
\end{aligned}
$$

By Theorem 5.4.1 and Cramér's theorem (Davidson (1994), Theorem 22.14), we have

$$
\begin{aligned}
\sqrt{T} y_n^m &= \frac{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+1)T2^{-(m+1)}-1} Z_t - \sum_{t=(\gamma_{m,n}+1)T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} Z_t}{b_T} \\
&\times (-1)^{\alpha_{m,n}} \frac{b_T\sqrt{T}}{\sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} Z_t + \sum_{t=\gamma_{m,n}T2^{-(m+1)}}^{(\gamma_{m,n}+2)T2^{-(m+1)}-1} \mathbb{E}I_t} \\
&\xrightarrow{D} \left\{B_\eta((\gamma_{m,n}+2)2^{-m+1}) - 2B_\eta((\gamma_{m,n}+1)2^{-m+1}) + B_\eta(\gamma_{m,n}2^{-m+1})\right\} \\
&\times \frac{(-1)^{\alpha_{m,n}} 2^{1/2} \left(\sum_{\tau=-\infty}^{\infty} \int_0^1 (\beta_j^\tau(z))^2 dz\right)^{1/2}}{\int_{\gamma_{m,n}2^{-m+1}}^{(\gamma_{m,n}+2)2^{-m+1}} \beta_j(z) dz},
\end{aligned}
$$

as $T \to \infty$. Denote the distributional limit by $\tilde{y}_n^m$. Set $Y_n^M = \sum_{m=0}^{M-1} y_n^m$ and $\tilde{Y}_n^M = \sum_{m=0}^{M-1} \tilde{y}_n^m$. Denote further $c_{(1)} = 2^{1/2}\left(\sum_{\tau=-\infty}^{\infty} \int_0^1 (\beta_j^\tau(z))^2 dz\right)^{1/2}$. We have

$$\sqrt{T} Y_n^M \xrightarrow{D} \tilde{Y}_n^M = c_{(1)} \sum_{m=0}^{M-1} (-1)^{\alpha_{m,n}} \times \\ \frac{B_\eta((\gamma_{m,n}+2)2^{-m+1}) - 2B_\eta((\gamma_{m,n}+1)2^{-m+1}) + B_\eta(\gamma_{m,n}2^{-m+1})}{\int_{\gamma_{m,n}2^{-m+1}}^{(\gamma_{m,n}+2)2^{-m+1}} \beta_j(z)dz},$$

as $T \to \infty$. It is immediate that $\mathbb{E}\tilde{Y}_n^M = 0$. We now look at the variance-covariance matrix of $\tilde{\mathbf{Y}}^M$. We have

$$\mathrm{Var}(\tilde{Y}_n^M) = \\ c_{(1)}^2 \left(\sum_{m=0}^{M-1} \frac{\eta((\gamma_{m,n}+2)2^{-(m+1)}) - \eta(\gamma_{m,n}2^{-(m+1)})}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} + 2\sum_{m=0}^{M-1}\sum_{m'=m+1}^{M-1} (-1)^{\alpha}_{m,n} \times \right. \\ (-1)^{\alpha_{m',n}} \left\{ \frac{-2\eta(\rho_{m,n} \wedge \gamma_{m',n}2^{-(m'+1)}) + 4\eta(\rho_{m,n} \wedge (\gamma_{m',n}+1)2^{-(m'+1)})}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz \int_{\gamma_{m',n}2^{-(m'+1)}}^{(\gamma_{m',n}+2)2^{-(m'+1)}} \beta_j(z)dz} + \right. \\ \frac{-2\eta(\rho_{m,n} \wedge (\gamma_{m',n}+2)2^{-(m'+1)}) + \eta(\gamma_{m',n}2^{-(m'+1)}) - 2\eta((\gamma_{m',n}+1)2^{-(m'+1)})}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz \int_{\gamma_{m',n}2^{-(m'+1)}}^{(\gamma_{m',n}+2)2^{-(m'+1)}} \beta_j(z)dz} \\ \left.\left. + \frac{\eta((\gamma_{m',n}+2)2^{-(m'+1)})}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz \int_{\gamma_{m',n}2^{-(m'+1)}}^{(\gamma_{m',n}+2)2^{-(m'+1)}} \beta_j(z)dz} \right\} \right),$$

where $\rho_{m,n} = (\gamma_{m,n}+1)2^{-(m+1)}$.

Diagonal contribution. Let us first consider the diagonal contribution to $\mathrm{Var}(\tilde{Y}_n^M)$. We have

$$\frac{c_{(1)}^2}{2} \frac{\eta((\gamma_{m,n}+2)2^{-(m+1)}) - \eta(\gamma_{m,n}2^{-(m+1)})}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} = \frac{\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta_j^\tau(z))^2 dz}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2}. \tag{5.29}$$

By Cauchy inequality and the extended mean-value theorem, we have

$$\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2 \leq 2^{-m}\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta_j(z))^2 dz =$$

$$2^{-m}\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \sum_\tau (\beta_j^\tau(z))^2 dz \frac{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta_j(z))^2 dz}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \sum_\tau (\beta_j^\tau(z))^2 dz} =$$

$$2^{-m}\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \sum_\tau (\beta_j^\tau(z))^2 dz \frac{(\beta_j(\omega))^2}{\sum_\tau (\beta_j^\tau(\omega))^2},$$

where $\omega \in [\gamma_{m,n}2^{-(m+1)}, (\gamma_{m,n}+2)2^{-(m+1)}]$. This, combined with (5.29), gives

$$2^{m+1}\inf_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} \leq c_{(1)}^2 \frac{\eta((\gamma_{m,n}+2)2^{-(m+1)}) - \eta(\gamma_{m,n}2^{-(m+1)})}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2}. \tag{5.30}$$

To obtain the upper bound, note that $\exists\, \omega_1, \omega_2 \in [\gamma_{m,n}2^{-(m+1)}, (\gamma_{m,n}+2)2^{-(m+1)}]$ such that

$$\sum_{\tau=-\infty}^{\infty}\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta_j^\tau(z))^2 dz \leq$$

$$2^m\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2 \frac{2^{-m}\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta_j(z))^2 dz}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} \times$$

$$\times \sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} =$$

$$2^m\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2 \frac{\beta_j(\omega_1)}{\beta_j(\omega_2)} \sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} \leq$$

$$2^m\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2 \left(1 + \frac{2^{-m}\sup_{\omega\in[0,1]}|\beta_j'(\omega)|}{\inf_{\omega\in[0,1]}\beta_j(\omega)}\right) \times$$

$$\times \sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2},$$

where β_j' is the one-sided derivative of β_j. The above, combined with (5.29), yields

$$c_{(1)}^2 \frac{\eta((\gamma_{m,n}+2)2^{-(m+1)}) - \eta(\gamma_{m,n}2^{-(m+1)})}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} \leq$$

$$2^{m+1}\sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} + \frac{2\sup_{\omega\in[0,1]}|\beta_j'(\omega)|}{\inf_{\omega\in[0,1]}\beta_j(\omega)} \sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} =$$

$$2^{m+1}\sup_{\omega\in[0,1]} \frac{\sum_\tau(\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} + O(1). \tag{5.31}$$

Off-diagonal contribution. Two cases are possible: either $\rho_{m,n} \geq (\gamma_{m',n}+2)2^{-(m'+1)}$ or $\rho_{m,n} \leq \gamma_{m',n}2^{-(m'+1)}$. In either of the two cases, we have

$$\begin{aligned}
&\Big|-2\eta\left(\rho_{m,n} \wedge \gamma_{m',n}2^{-(m'+1)}\right) + 4\eta\left(\rho_{m,n} \wedge (\gamma_{m',n}+1)2^{-(m'+1)}\right) + \\
&\quad -2\eta\left(\rho_{m,n} \wedge (\gamma_{m',n}+2)2^{-(m'+1)}\right) + \eta\left(\gamma_{m',n}2^{-(m'+1)}\right) - \\
&\quad 2\eta\left((\gamma_{m',n}+1)2^{-(m'+1)}\right) + \eta\left((\gamma_{m',n}+2)2^{-(m'+1)}\right)\Big| = \Big|\eta\left(\gamma_{m',n}2^{-(m'+1)}\right) - \\
&\quad 2\eta\left((\gamma_{m',n}+1)2^{-(m'+1)}\right) + \eta\left((\gamma_{m',n}+2)2^{-(m'+1)}\right)\Big| \leq 2^{-2m'-1} \sup_{\omega\in[0,1]} |\eta''(\omega)|,
\end{aligned}$$

where the last inequality follows by the mean-value theorem and η'' denotes the one-sided derivative of η'. Using the above, and, again, the mean-value theorem, we bound the off-diagonal contribution by

$$\begin{aligned}
&2c_{(1)}^2 \sum_{m=0}^{M-1} \frac{1}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz} \sum_{m'=m+1}^{M-1} 2^{-(m'+1)} \frac{2^{-m'} \sup_{\omega\in[0,1]} |\eta''(\omega)|}{\int_{\gamma_{m',n}2^{-(m'+1)}}^{(\gamma_{m',n}+2)2^{-(m'+1)}} \beta_j(z)dz} \leq \\
&\quad \frac{2c_{(1)}^2 \sup_{\omega\in[0,1]} |\eta''(\omega)|}{\inf_{\omega\in[0,1]} \beta_j(\omega)} \sum_{m=0}^{M-1} \frac{1}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz} \sum_{m'=m+1}^{M-1} 2^{-(m'+1)} \leq \\
&\quad \frac{2c_{(1)}^2 \sup_{\omega\in[0,1]} |\eta''(\omega)|}{\inf_{\omega\in[0,1]} \beta_j(\omega)} \sum_{m=0}^{M-1} \frac{2^{-(m+1)}}{\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz} \leq \\
&\quad \frac{c_{(1)}^2 \sup_{\omega\in[0,1]} |\eta''(\omega)|}{\inf_{\omega\in[0,1]} (\beta_j(\omega))^2} M = O(M). \qquad (5.32)
\end{aligned}$$

Putting together (5.30), (5.31) and (5.32), we finally arrive at

$$\begin{aligned}
&\left(2^{M+1}-2\right) \inf_{\omega\in[0,1]} \frac{\sum_\tau (\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} - O(M) \leq \mathrm{Var}(\tilde{Y}_n^M) \\
&\quad \leq \left(2^{M+1}-2\right) \sup_{\omega\in[0,1]} \frac{\sum_\tau (\beta_j^\tau(\omega))^2}{(\beta_j(\omega))^2} + O(M). \qquad (5.33)
\end{aligned}$$

Let us now consider $\mathrm{Cov}(\tilde{Y}_{n_1}^M, \tilde{Y}_{n_2}^M)$ for $n_1 \neq n_2$. Let $M' = \#\{m : \tilde{y}_{n_1}^m = \tilde{y}_{n_2}^m\}$. Let us look at the case $M' > 0$ (the case $M' = 0$ is straightforward). It is easy to show that

$$\begin{aligned}
&\mathrm{Cov}(\tilde{Y}^M_{n_1}, \tilde{Y}^M_{n_2}) = \\
&\quad \mathrm{Cov}\left(\sum_{m=0}^{M'-1} \tilde{y}^m_{n_1} + \tilde{y}^{M'}_{n_1} + \sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_1}, \sum_{m=0}^{M'-1} \tilde{y}^m_{n_1} - \tilde{y}^{M'}_{n_1} + \sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_2}\right) = \\
&\quad \mathrm{Var}\left(\sum_{m=0}^{M'-1} \tilde{y}^m_{n_1}\right) - \mathrm{Var}\left(\tilde{y}^{M'}_{n_1}\right) + \mathbb{E}\left(\sum_{m=0}^{M'-1} \tilde{y}^m_{n_1}\left(\sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_1} + \sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_2}\right) + \right. \\
&\quad \left. \tilde{y}^{M'}_{n_1}\left(\sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_1} - \sum_{m=M'+1}^{M-1} \tilde{y}^m_{n_2}\right)\right).
\end{aligned}$$

The expectation can be shown to be $O(M)$ using the same methodology as for bounding the off-diagonal component of $\mathrm{Var}(\tilde{Y}^M_n)$. We will now show that $\mathrm{Var}(\sum_{m=0}^{M'-1} \tilde{y}^m_{n_1}) - \mathrm{Var}(\tilde{y}^{M'}_{n_1}) = O(M)$. We first quote two simple facts: let g be a continuous function with a bounded one-sided derivative over $[0,1]$ and let $[c,d] \subset [a,b] \subset [0,1]$. We have

$$\left|\int_a^b g(z)dz - \frac{b-a}{d-c}\int_c^d g(z)dz\right| \leq (b-a)^2 \sup_z |g'(z)| \tag{5.34}$$

$$\left|\left(\int_c^d g(z)dz\right)^2 - \left(\frac{d-c}{b-a}\right)^2\left(\int_a^b g(z)dz\right)^2\right| \leq (d-c)^2(b-a) \sup_z |(g^2(z))'| \tag{5.35}$$

For simplicity, denote $n = n_1$. Using again the same method as for bounding the off-diagonal component of the variance, we obtain

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{m=0}^{M'-1} \tilde{y}^m_n\right) - \mathrm{Var}(\tilde{y}^{M'}_n) = O(M) + 2\sum_{m=0}^{M'-1} \frac{\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta^\tau_j(z))^2 dz}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} - \\
&\quad \frac{2\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} (\beta^\tau_j(z))^2 dz}{\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} \beta_j(z)dz\right)^2} = O(M) + 2\sum_{m=0}^{M'-1} \\
&\quad \left(\frac{\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} (\beta^\tau_j(z))^2 dz}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}} \beta_j(z)dz\right)^2} - \frac{\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} (\beta^\tau_j(z))^2 dz}{2^{M'-m}\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} \beta_j(z)dz\right)^2}\right) \\
&\quad + \frac{2^{-M'+1}\sum_{\tau=-\infty}^{\infty} \int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} (\beta^\tau_j(z))^2 dz}{\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}} \beta_j(z)dz\right)^2}.
\end{aligned}$$

Consider a single component of the sum over m: it is a difference of two ratios which we denote here by $I - II$ to shorten the notation. We have $|I - II| \leq |I - III| + |III - II|$, where

$$III = \frac{2^{M'-m}\sum_{\tau=-\infty}^{\infty}\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}(\beta_j^{\tau}(z))^2 dz}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}}\beta_j(z)dz\right)^2}.$$

Using (5.34), we get

$$|I - III| \leq \frac{2^{-2m}\sup_{\omega\in[0,1]}\left|\left(\sum_{\tau}(\beta_j^{\tau}(\omega))^2\right)'\right|}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}}\beta_j(z)dz\right)^2} \leq \frac{\sup_{\omega\in[0,1]}\left|\left(\sum_{\tau}(\beta_j^{\tau}(\omega))^2\right)'\right|}{\left(\inf_{\omega\in[0,1]}\beta_j(\omega)\right)^2}.$$

On the other hand, using (5.35) we have

$$|III - II| = 2^{M'-m}\left|\frac{\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}\beta_j(z)dz\right)^2 - 2^{2(m-M')}\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}}\beta_j(z)dz\right)^2}{\left(\int_{\gamma_{m,n}2^{-(m+1)}}^{(\gamma_{m,n}+2)2^{-(m+1)}}\beta_j(z)dz\right)^2}\right|$$
$$\times\frac{\sum_{\tau=-\infty}^{\infty}\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}(\beta_j^{\tau}(z))^2 dz}{\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}\beta_j(z)dz\right)^2} \leq$$
$$\frac{2^{-M'}\sum_{\tau=-\infty}^{\infty}\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}(\beta_j^{\tau}(z))^2 dz}{\left(\int_{\gamma_{M',n}2^{-(M'+1)}}^{(\gamma_{M',n}+2)2^{-(M'+1)}}\beta_j(z)dz\right)^2}\frac{\sup_{\omega\in[0,1]}((\beta_j(\omega))^2)'}{\left(\inf_{\omega\in[0,1]}\beta_j(\omega)\right)^2},$$

which is bounded by (5.33). This proves the assertion that $\mathrm{Var}(\sum_{m=0}^{M'-1}\tilde{y}_{n_1}^m) - \mathrm{Var}(\tilde{y}_{n_1}^{M'}) = O(M)$. Setting $V_n^M = \sum_{m=0}^{M-1} v_n^m$ completes the proof of the theorem. □

5.5.2 Properties of the Haar-Fisz transform for $M = \log_2(T)$

In the asymptotic framework set out in Section 5.5.1, we assume that M is fixed, and therefore the length of the Haar-Fisz-transformed vector $\mathcal{F}^M\mathbf{I}$ is always constant and equal to 2^M, even though $T \to \infty$. This ensures the asymptotic Gaussianity of $\mathcal{F}^M\mathbf{I}$, in the sense specified by Theorem 5.5.1. However, to obtain an

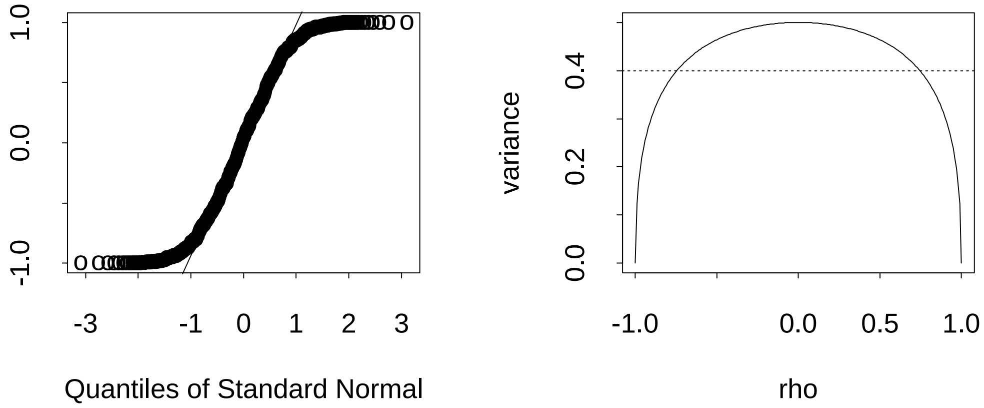

Figure 5.3: Left plot: the q-q plot of $\mathbf{f}^9$ arising from the Haar periodogram of a pure white noise process at scale $j = -1$ (against the normal quantiles). Right plot: solid line — the variance of $f_n^{\log_2(T)-1}$ against the correlation of the Gaussian variables involved; dotted line — variance = 0.4 (see text for further description).

invertible operator, we need to set $M = \log_2(T)$. Simulations suggest that the asymptotic distribution of $\mathcal{F}^{\log_2(T)}$ is not exactly Gaussian, which is not surprising given the fact that the distribution of

$$f_n^{\log_2(T)-1} = \frac{I_{2n,T}^{(j)} - I_{2n+1,T}^{(j)}}{I_{2n,T}^{(j)} + I_{2n+1,T}^{(j)}}$$

(see the second example in Section 5.3.2) is far from Gaussian. To illustrate this statement, let us consider the Haar periodogram sequence $I_{t,1024}^{(-1)}$ of a pure white noise process. The left plot in Figure 5.3 shows the q-q plot of the corresponding sequence $\mathbf{f}^9$ against the normal quantiles: its distribution is strongly deviated from Gaussian in the tails. Other extensive simulations have shown that, for a wide range of processes, the distribution of $\mathbf{f}^M$ gets closer to Gaussianity as M decreases (as expected, see proof of Theorem 5.5.1).

However, even though taking $M = \log_2(T)$ (instead of keeping it fixed) spoils the asymptotic Gaussianity of $\mathcal{F}^M\mathbf{I}$, it does not seem to upset the other important property of $\mathcal{F}^M\mathbf{I}$: the variance stabilisation. To illustrate this point, we consider the variance of the summand $f_n^{\log_2(T)-1}$. Note that $f_n^{\log_2(T)-1}$ is always of the form $f_n^{\log_2(T)-1} = (\zeta_1^2 - \zeta_2^2)/(\zeta_1^2 + \zeta_2^2)$, where (ζ_1, ζ_2) is bivariate normal with mean $(0, 0)$. For simplicity, we assume that $\mathrm{Var}(\zeta_1) = \mathrm{Var}(\zeta_2)$, which is not a restrictive assumption: due to the local stationarity property, the two variances tend to the same limit as $T \to \infty$. Let $\rho = \mathrm{corr}(\zeta_1, \zeta_2)$. By straightforward computation, it can be shown that

$$\mathrm{Var}\left(f_n^{\log_2(T)-1}\right) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\left((1-u^2)^2 - \left((1-u^2)\,\rho + 2u\sqrt{1-\rho^2}\right)^2\right)^2}{(1+u^2)\left((1-u^2)^2 + \left((1-u^2)\,\rho + 2u\sqrt{1-\rho^2}\right)^2\right)^2} du.$$

The right plot in Figure 5.3 shows the graph of $\mathrm{Var}\left(f_n^{\log_2(T)-1}\right)$ against ρ. It can be seen that $\mathrm{Var}\left(f_n^{\log_2(T)-1}\right)$ is "stable" for a wide range of correlation values: indeed, the variance is between 0.4 and 0.5 for $|\rho| \leq 0.74$. This implies that while incorporating $\mathbf{f}^{\log_2(T)-1}$ spoils the asymptotic Gaussianity property of the Haar-Fisz transform, it helps achieve its variance stabilisation property.

A similar variance stabilisation phenomenon occurs for $\mathbf{f}^M$ for $M < \log_2(T) - 1$.

5.5.3 Simulation

As an illustration of the Gaussianisation and the variance stabilisation properties of the Haar-Fisz transform, consider the process $X_{t,T} = \sigma(t/T)Y_{t,T}$, where $Y_{t,T} = \rho(t/T)Y_{t-1,T} + \varepsilon_t$ with $|\rho(z)| < 1$ and $\varepsilon_t \sim N(0, 1)$ i.i.d. It can easily be shown that the local autocovariance function for $X_{t,T}$ has the form

$$c(z, \tau) = \sigma^2(z)\frac{\rho(z)^\tau}{1 - \rho(z)^2}$$

and, for $\beta_j(z)$ arising from the Haar periodogram, we have

$$\beta_j(z) = \sigma^2(z)\frac{1 - \rho(z)^2 + 2^{j+3}\rho(z)^{2^{-j-1}+1} - 6\,2^j\rho(z) - 2^{j+1}\rho(z)^{2^{-j}+1}}{(1-\rho(z)^2)(1-\rho(z))^2}.$$

We consider the following two cases:

TVAR. $\sigma^2(z) = 1$ and $\rho(z) = 1.8z - 0.9$, so that $X_{t,T}$ is a time-varying AR(1) process;

TMWN. $\sigma^2(z)$ is a scaled Donoho & Johnstone (1995) `bumps` function with (min, max) values of (1/8, 8), and $\rho(z) = 0$, so that $X_{t,T}$ is a time-modulated white noise process.

In both of these models, we simulate 100 sample paths for both $T = 256$ and $T = 1024$. For each of the simulated sample paths, we compute the wavelet periodogram at scales $j = -1, \ldots, -\log_2(T)$. For each of the periodogram sequences $I_{t,T}^{(j)}$ obtained in this way, we compute the residuals $\mathcal{F}^M I_{t,T}^{(j)} - \mathcal{F}^M \beta_j(t/T)$ for $M = \log_2(T) - 2, \log_2(T) - 1, \log_2(T)$. We assess the Gaussianity of each sequence of residuals by looking at the p-value of the Kolmogorov-Smirnov statistic, returned by the S-Plus function `ks.gof`. For comparison, we also consider the residuals from the log transform: $\log(I_{t,T}^{(j)}) - \log(\beta_j(t/T))$.

The results of the experiment are shown in Figure 5.4. We observe that for $M = \log_2(T) - 2$, the proportion of p-values exceeding 5% is close to 95% for $j = -1, \ldots, -5$, so that residual sequences at these scales can be regarded as approximately Gaussian. However, even for $M = \log_2(T)$ the proportion of p-values exceeding 5% is incomparably larger than the same proportion computed for the log transform. Indeed, for $T = 1024$, no p-value exceeded the 5% threshold for the log transform.

The above experiment demonstrates that even for $M = \log_2(T)$ (the invertible case), the Haar-Fisz transform is a far better Gaussianiser than the log transform.

In practice, we often observe a degree of correlation in $\mathcal{F}^M \mathbf{I}^{(j)}$, particularly at coarser scales, i.e. for large negative j. This has to be taken into account when denoising Haar-Fisz transformed sequences.

5.6 Denoising the wavelet periodogram

In this section, we first outline our general methodology for denoising the wavelet periodogram of a Gaussian LSW process $X_{t,T}$, basing on a single stretch of observations. Then, we provide simulation results which demonstrate the effectiveness

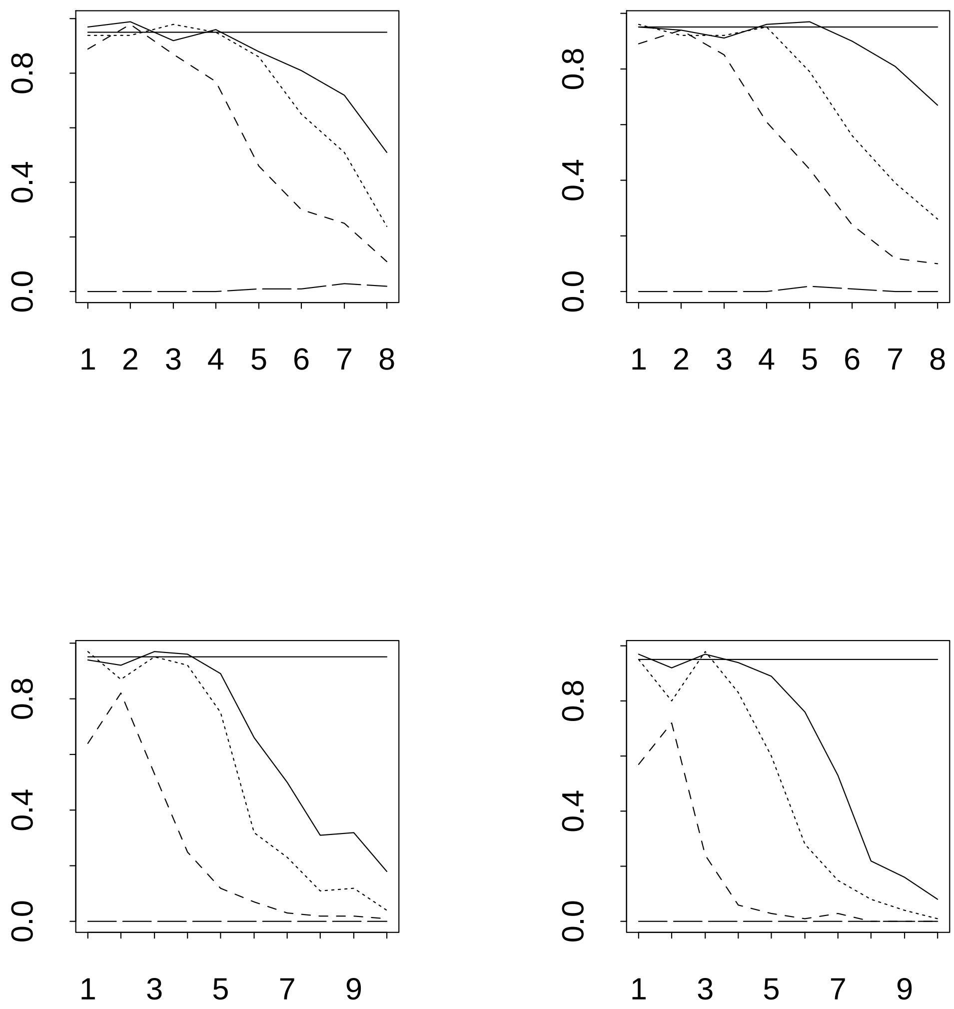

Figure 5.4: Proportion of p-values exceeding or equal to 5% (x-axis shows negative scale $-j$). Left column: results for TVAR, right column: results for TMWN. Top row: $T = 256$, Bottom row: $T = 1024$. Solid line: $M = \log_2(T)$, dotted line: $M = \log_2(T) - 1$, dashed line: $M = \log_2(T) - 2$, long-dashed line: the log transform. Horizontal solid line: 0.95.

of our technique.

The generic algorithm consists of the following steps.

1. For each $j = -1, \dots, -J(T)$, compute the raw wavelet periodogram $I_{t,T}^{(j)}$. In practice, this is done by taking the non-decimated wavelet transform of $X_{t,T}$ down to the level $-J(T)$, and then squaring the result. For computational convenience, we use periodic boundary treatment; another option would be to use e.g. symmetric boundary treatment.

2. For each $j = -1, \dots, -J(T)$, take the Haar-Fisz transform of $I_{t,T}^{(j)}$ at a fixed resolution level $M \leq \log_2(T)$.

3. For each j, denoise the Haar-Fisz transformed periodogram sequence using any wavelet denoising technique suitable for correlated Gaussian noise with constant variance. The wavelet denoising procedure employed at this stage may be of a translation-invariant (TI) type: we refer to TI-denoising at this stage as "internal" cycle-spinning (CS).

4. For each j, take the inverse Haar-Fisz transform of the denoised data.

5. If $M < \log_2(T)$, then for each j interpolate the estimates obtained in this way to the grid $\{t/T\}_{t=0}^{T-1}$ (so that they are of length T and not $2^M < T$). In our empirical investigation, we used simple linear interpolation. For each j, take the result to be an estimate of $\beta_j(z)$.

6. For a fixed integer S, let $s = 1, \dots, S - 1$. For each j, shift $I_{t,T}^{(j)}$ cyclically by s, denoise the shifted version using steps 2. – 5. of this algorithm, and shift back by s to obtain an estimate of $\beta_j(z)$. The CS at this stage is referred to as "external" cycle-spinning.

7. For each j, the *final estimate* of $\beta_j(z)$ is obtained by averaging over the estimates obtained through the S shifts.

A few remarks are in order.

Computational complexity. Steps 1. – 5. of the algorithm are each of computational order $O(TJ(T))$, provided that the wavelet denoising method used in step 3. has complexity $O(T)$. Therefore, the whole algorithm 1. – 7. is of computational order $O(STJ(T))$. In practice, the software is fast.

Use of wavelets. It is worth recalling here that, effectively, we use wavelets at four different stages of the denoising procedure:

1. First of all, a non-decimated wavelet system ψ is used in the construction of the LSW process $X_{t,T}$.
2. The same system ψ is used to compute the wavelet periodogram $I_{t,T}^{(j)}$ in step 1. of the denoising algorithm.
3. The (inverse) Haar-Fisz transform in step 2. (4.) relies on the Haar transform: thus, wavelets are used for the third time.
4. Finally, we use wavelets (possibly a different family, say $\tilde{\psi}$) to denoise the Haar-Fisz transformed periodogram in step 3.

Cycle-spinning. Let $\mathcal{S}$ be the shift-by-one-operator from Nason & Silverman (1995). The Haar-Fisz transform is not translation-equivariant since $\mathcal{S}\mathcal{F}^M \neq \mathcal{F}^M\mathcal{S}$. Therefore, it is potentially beneficial to apply the external CS of step 6. even if step 3. uses internal CS.

We now move on to describe our particular simulation setup.

5.6.1 Simulation

In this section, we describe the details of our simulation study which compares the performance of our Haar-Fisz denoising algorithm with the original technique of Nason *et al.* (2000).

The "test processes" used in this section are the same as those in Section 5.5.3: TVAR and TMWN. We consider the Haar periodogram of TVAR and TMWN, for sample paths of length 256 and 1024. In step 3. of the Haar-Fisz denoising algorithm, we use non-TI level-dependent universal hard thresholding, appropriate for correlated Gaussian data as described in Johnstone & Silverman (1997). At this stage, we use Daubechies' Least Asymmetric wavelets with 4 vanishing moments, in both our algorithm and that of Nason *et al.* (2000).

Computational experiments suggest that for correlated noise, the choice of primary resolution (PR) is of utmost importance. We do not choose the PR automatically (actually, we are unaware of any existing technique for performing automatic PR selection when the noise is correlated), but instead, we subjectively choose the PR for which the method of Nason *et al.* (2000) gives the most visually appealing results for the wavelet periodogram at the finest scale, i.e. $j = -1$. We also use the same PR in our algorithm. The particular values of the PR are:

- 7 for TMWN 1024;
- 6 for TMWN 256;
- 4 for TVAR 1024;
- 3 for TVAR 256.

We use $S = 10$ external cycle-shifts. Using more shifts is likely to be beneficial in terms of MISE but is also more burdensome computationally. We only report results for $M = \log_2(T)$ (i.e. for the full invertible Haar-Fisz transform).

Figure 5.5 shows estimates of the local variance constructed from the estimates of the periodogram (formula (3.65)) obtained using the two methods described above, for particular sample paths of TMWN 1024 and TVAR 1024. For both sample paths, our method achieves lower ISE.

Figure 5.6 shows, for each j, the differences between the logarithm of the ISE in estimating $\beta_j(z)$ for the method of Nason *et al.* (2000), and for our Haar-Fisz algorithm. The results are averaged over 100 simulated sample paths. Our algorithm is superior in most of the cases, except for the 4 finest scales in TVAR

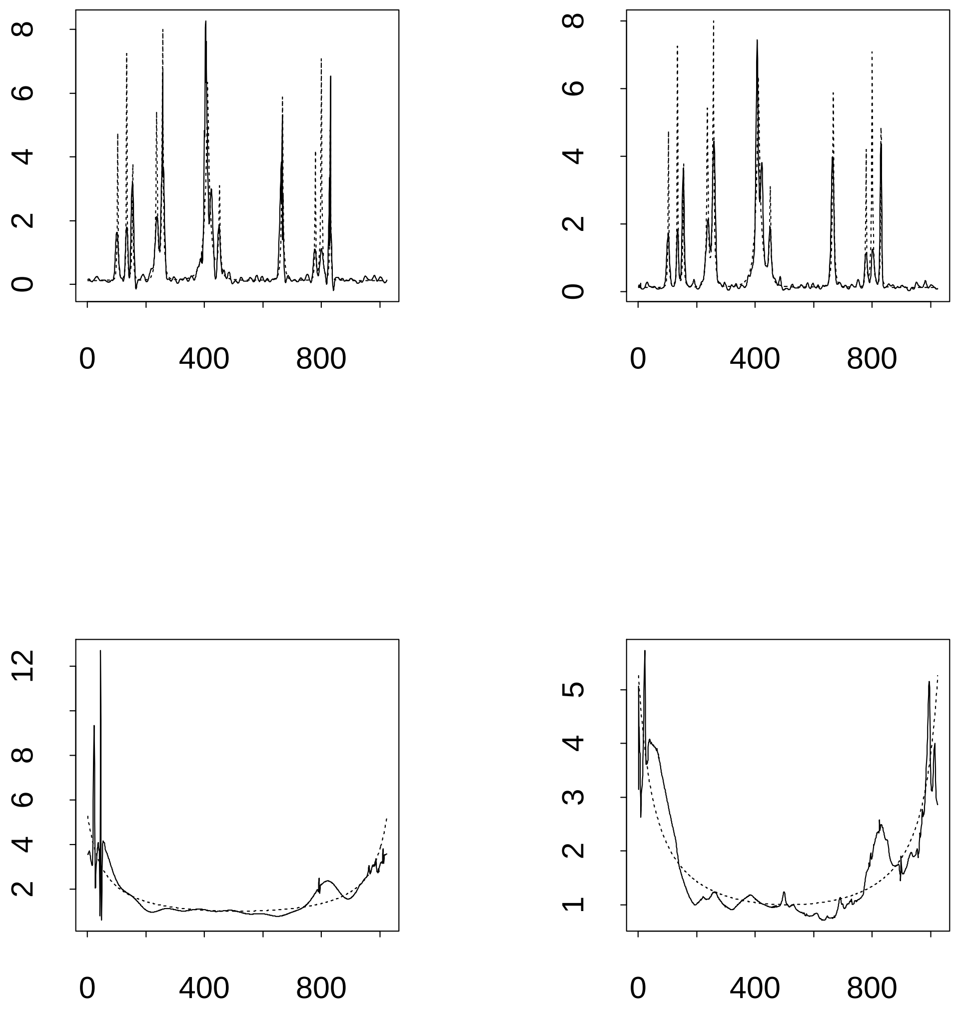

Figure 5.5: Solid lines: estimates of the local variances for $T = 1024$ in the TMWN model (top row), and the TVAR model (bottom row), using the method of Nason *et al.* (2000) (left column) and the Haar-Fisz algorithm (right column) as described in the text. Dotted lines: true local variances.

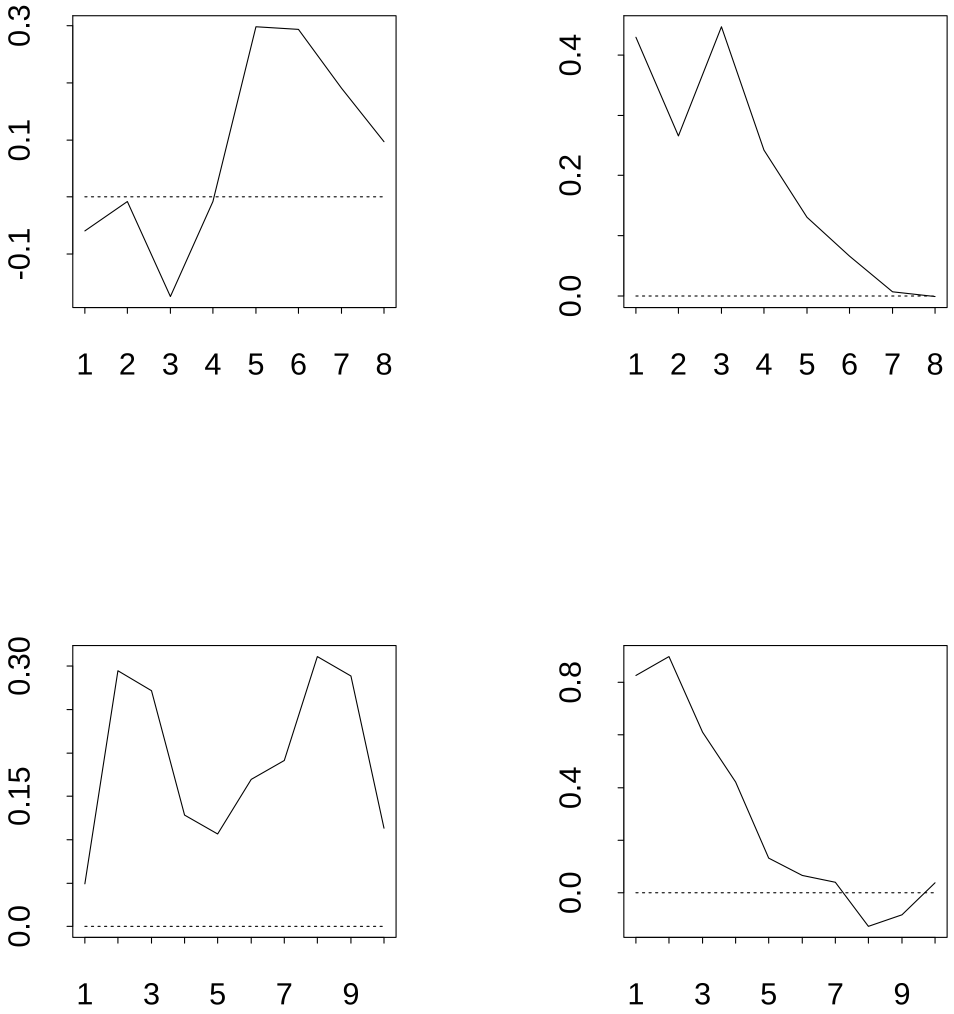

Figure 5.6: Solid line: difference between logged MISE for Nason *et al.* (2000) and for our Haar-Fisz algorithm (x-axis shows negative scale $-j$). Positive value means our algorithm does better. Left column: results for TVAR, right column: results for TMWN. Top row: $T = 256$, bottom row: $T = 1024$. Dotted line: zero.

256, and the 3 coarsest scales in TMWN 1024. A similar pattern has been obtained for other values of the PR.

We have also performed additional simulations for $M = \log_2(T) - 1$ and $M = \log_2(T) - 2$. It turned out that as long as the PR remained fixed, the choice of M had very little influence upon the estimates.

On a final note, it must be mentioned that other denoising methods can also be used in step 3., and our algorithm can only benefit from this flexibility. Some of the techniques for correlated data are reviewed in Opsomer *et al.* (2001). We have also experimented with the eBayes method of Johnstone & Silverman (2003) and obtained encouraging results.

5.7 Real data example: the Dow Jones index

In this section, we perform a local variance analysis of the DJIA series $D_{t,T}$ of Section 4.5 ($T = 1024$). The source of the data is `http://bossa.pl/notowania/daneatech/metastock` (page in Polish). We used the following four methods to compute the local variance of $D_{t,T}$:

1. Our Haar-Fisz method of Section 5.6, based on the Haar periodogram, with the following parameters: $M = 10$, $S = 10$, step 3. applied non-TI level-dependent hard universal thresholding using Daubechies' Least Asymmetric wavelet with 4 vanishing moments. PR = 7.

2. Our Haar-Fisz method of Section 5.6, based on the Haar periodogram, with the following parameters: $M = 10$, $S = 10$, step 3. used the S-Plus spline smoothing routine `smooth.spline` with default parameters.

3. A modification of our Haar-Fisz method: instead of the sequences of the wavelet periodogram of $D_{t,T}$, the input to the Haar-Fisz algorithm was $D^2_{t,T}$. We took the smoothed version of $D^2_{t,T}$ to be the estimate of the local variance. The parameters of the Haar-Fisz algorithm were: $M = 10$, $S = 10$, step

3. used the S-Plus spline smoothing routine `smooth.spline` with default parameters.

4. The method of Nason *et al.* (2000) with the following parameters: TI level-dependent universal hard thresholding using Daubechies' Least Asymmetric wavelet with 4 vanishing moments, PR = 7. The `smooth.dev` parameter in the `ewspec` routine (Nason (1998)) was set to `var`.

The results for PR $\neq$ 7 were less convincing. Figure 5.7 shows all four estimates plotted on a log scale. The two estimates based on spline smoothing show the least variability, the estimate 4. is the most variable, and the estimate 1. — the second most variable. Moreover, 1. estimates the variance at a slightly higher level than the other three methods.

One interesting question which can be asked is whether or not $D_{t,T}$ can be modelled as Gaussian. This can be examined, for example, by dividing $D_{t,T}$ by the square root of the estimates of the local variance, and looking at the distribution of the residuals. Figure 5.8 shows the `qqnorm` plot of the empirical quantiles of the residuals against the quantiles of the standard normal, for the four methods described above. The surprising observation is that all four plots consistently indicate that the upper tail is slightly platykurtic. However, there is no consistency in the assessment of the behaviour of the lower tail: here, 3 plots indicate platykurtosis, but the result of method 3. suggests slight leptokurtosis.

However, the p-values of the Kolmogorov-Smirnov test (returned by the S-Plus routine `ks.gof`) are large for each of the 4 sequences of residuals. In this sense, it can be concluded that the departure of $D_{t,T}$ from Gaussianity is insignificant.

This is in stark contrast to stationary nonlinear modelling (e.g. (G)ARCH or Stochastic Volatility), where, typically, the marginal distribution of financial log-returns is modelled as heavily leptokurtic.

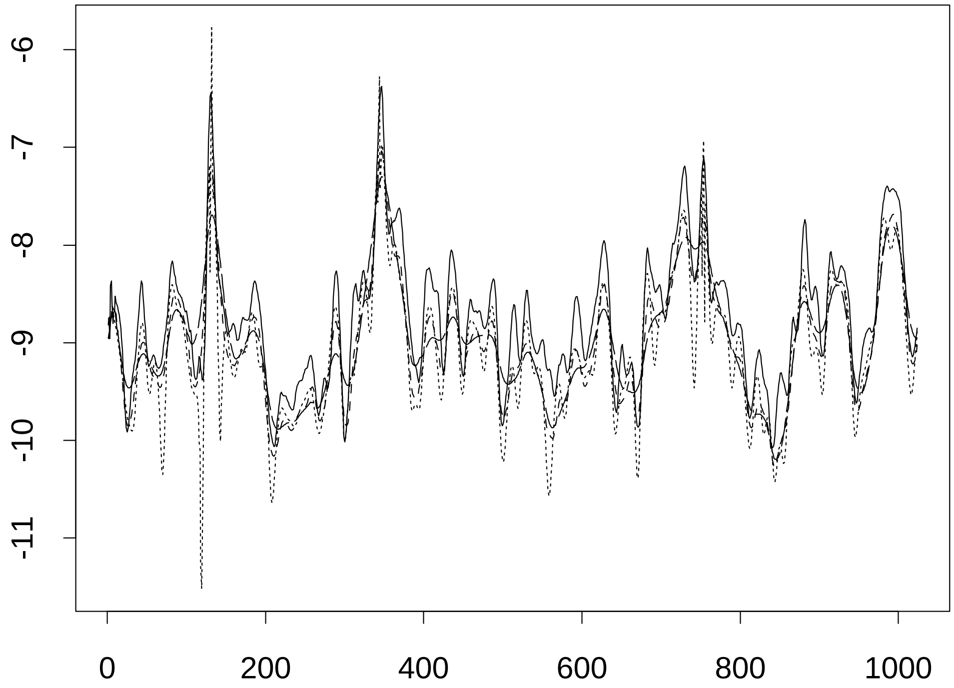

Figure 5.7: Four estimates of the local variance of $D_{t,T}$ on a log scale. Solid line: method 1. Dashed line: method 2. Long-dashed line: method 3. Dotted line: method 4. See text for further description.

5.8 Conclusion

In this chapter, we have introduced a Haar-Fisz variance-stabilising transform for the wavelet periodogram (WP) of a Gaussian LSW process. The transform, performed in the wavelet domain by dividing the Haar detail coefficients of the WP by the corresponding smooth coefficients (an instance of the so-called Fisz transform), brings the distribution of the WP closer to normality, as well as stabilising its variance. This makes the WP more amenable to standard denoising techniques which require stationary Gaussian noise. The computational complexity of the Haar-Fisz transform is linear in the number of data points, which is required to

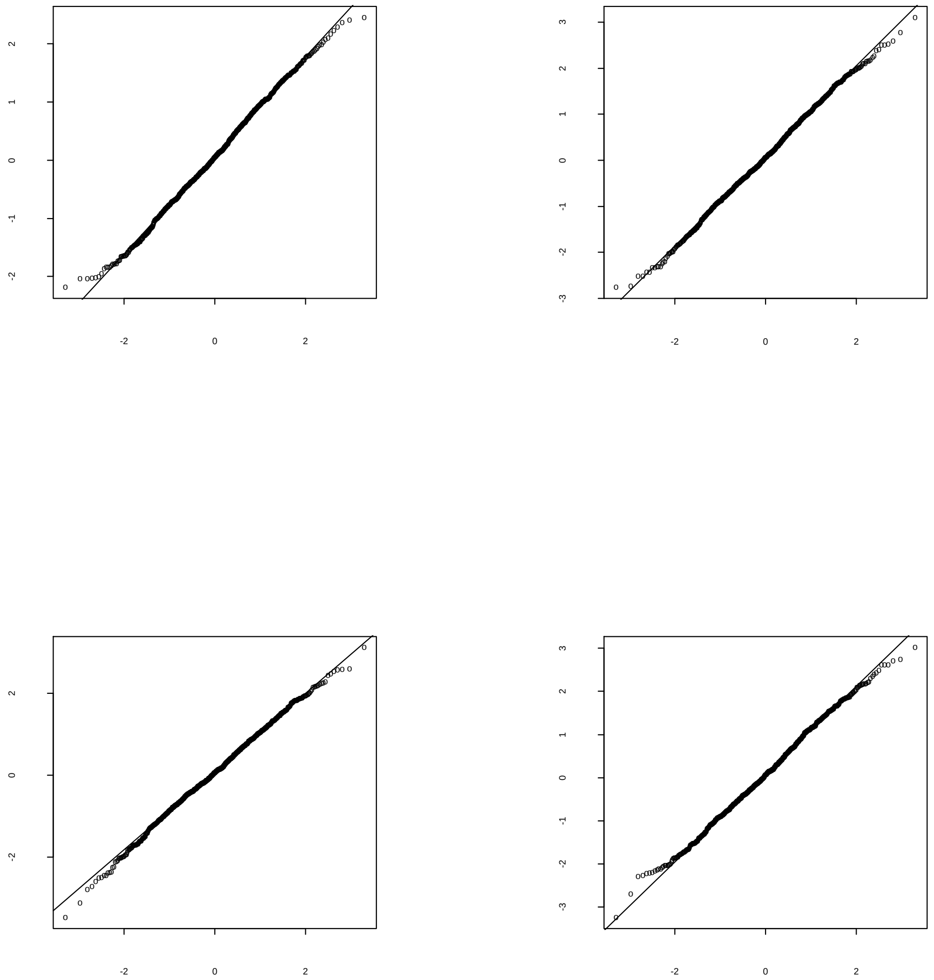

Figure 5.8: Empirical quantiles of the residuals of $D_{t,T}$ against the quantiles of the standard normal. Top left: method 1. Top right: method 2. Bottom left: method 3. Bottom right: method 4. See text for further description.

be a power of two.

In order to analyse theoretical properties of the Haar-Fisz transform in a certain asymptotic setting, we have formulated and proved a functional central limit theorem (FCLT) for the centred WP. Next, we have applied our FCLT to demonstrate the Gaussianising, variance-stabilising and decorrelating properties of the Haar-Fisz transform in the case where the length of the output vector remains constant as the length of the input vector goes to infinity.

Exact asymptotic Gaussianity does not hold if the length of the output vector of the Haar-Fisz transform matches the length of the input vector (which is the more interesting case in practice). However, we have provided some numerical evidence that the limiting distribution is still not far from Gaussian, and that its variance is well stabilised. Extensive simulations have shown that even in this case, the Haar-Fisz transform is a far more effective Gaussianiser than the usual log transform.

Next, we considered a denoising algorithm for the WP, based on the Haar-Fisz transform. Theory has shown that the new algorithm is computationally fast, and simulation — that its MISE performance is better than that of the existing competitor.

Finally, several variants of the algorithm have been used to compute the local variance of the time series of daily log returns on the Dow Jones index. All of them consistently demonstrated that the series can be modelled as Gaussian.

The S-Plus routines implementing the algorithm, as well as the data set, are included on the associated CD.

Chapter 6

A Haar-Fisz algorithm for Poisson intensity estimation

In this chapter, we propose a Haar-Fisz-type algorithm for estimating the discretised intensity function of an inhomogeneous one-dimensional Poisson process, in the regression setting specified in Section 2.3.3. The Haar-Fisz principle was already introduced in Chapter 5: take the Haar transform of the data, divide the arising detail coefficients by the corresponding smooth coefficients raised to an appropriate power, and then take the inverse Haar transform. In this chapter, we apply this algorithm to sequences of Poisson counts, with the aim of stabilising their variance and bringing their distribution close to Gaussianity. This then enables us to apply known denoising techniques suitable for i.i.d. Gaussian noise to estimate the underlying Poisson intensity. Simulations demonstrate that our denoising method usually significantly outperforms the existing state-of-the-art techniques.

Some results of this chapter were used, in a modified form, in the article by P. Fryzlewicz and G. P. Nason (2003) "A Haar-Fisz algorithm for Poisson intensity estimation", to appear in the Journal of Computational and Graphical Statistics. Throughout the thesis, this article will be referred to as Fryzlewicz & Nason (2004).

6.1 The Fisz transform for Poisson variables

In this section, we come back to the theorem by Fisz quoted in Section 5.1 as Theorem 5.1.1, and attempt to apply it to independent Poisson variables. Let $\xi(\lambda)$ be a Poisson variable with mean λ. It is shown in the original paper by Fisz (1955) that $\xi(\lambda)$ satisfies the first two assumptions of Theorem 5.1.1. Using the original notation from Fisz (1955), assume that $\xi_1(\lambda_1)$, $\xi_2(\lambda_2)$ are independent and $\lambda_1/\lambda_2 \to 1$ as $(\lambda_1, \lambda_2) \to (\infty, \infty)$. Then, $\zeta(\lambda_1, \lambda_2)$ (see formula (5.2)) is asymptotically normal

$$N\left(\frac{\lambda_2 - \lambda_1}{(\lambda_2 + \lambda_1)^p}, \frac{\sqrt{\lambda_2 + \lambda_1}}{(\lambda_2 + \lambda_1)^p}\right).$$

Note that setting $p = 1/2$ makes the variance of $\zeta(\lambda_1, \lambda_2)$ independent of λ. Therefore, to achieve the variance stabilisation property of the Haar-Fisz transform for Poisson data, we shall need to divide the Haar detail coefficients by the *square root* of the corresponding smooth coefficients. This is in contrast to the wavelet periodogram case where power $p = 1$ had to be used.

Recall from Section 5.1 the definition of $\zeta_{1/2}$ (the Fisz transform with exponent 1/2):

$$\zeta_{1/2}(X_1, X_2) = (X_1 - X_2)/(X_1 + X_2)^{1/2}, \tag{6.1}$$

with the convention $0/0 = 0$. Let $X_i \sim \text{Pois}(\lambda_i)$ for $i = 1, 2$ and X_1, X_2 independent. Theorem 5.1.1 only determines the behaviour of $\zeta_{1/2}(X_1, X_2)$ if $(\lambda_1, \lambda_2) \to (\infty, \infty)$ and $\lambda_1/\lambda_2 \to 1$. In practice, it would be useful to investigate the behaviour of $\zeta_{1/2}(X_1, X_2)$ if the means of X_1, X_2 are not necessarily "large" or "close". More specifically, we are interested in

- how well the Fisz transform can Gaussianise and stabilise variance,
- how well we can determine the mean, i.e. how close $\mathbb{E}\{\zeta_{1/2}(X_1, X_2)\}$ is to $\zeta_{1/2}(\lambda_1, \lambda_2)$,

for a whole range of λ_i. These issues would be challenging to investigate theoretically. However, to cast some light we performed the following simulation experiment. We chose values of λ_i to range from 1 to 40 in steps of 1. For each pair (λ_1, λ_2) we drew 10^5 values of $\zeta_{1/2}(X_1, X_2)$ as defined by (6.1) and denoted the sample by $\mathbf{z}(\lambda_1, \lambda_2)$. For a comparison of Gaussianisation we also computed Anscombe's transform, as mentioned in Section 2.3.3, to the X_i which arose from the larger λ_i (this comparison was charitable to Anscombe: either X_1 or X_2 could be used but Anscombe works better for larger intensities).

Figure 6.1 gives some idea of how well the Fisz transform Gaussianises, stabilises variance and how close $\bar{\mathbf{z}}(\lambda_1, \lambda_2)$ is to $\zeta_{1/2}(\lambda_1, \lambda_2)$. The top left figure shows that Fisz is always "more Gaussian" than Anscombe. The top right figure merely shows that $\bar{\mathbf{z}}(\lambda_1, \lambda_2)$ is very close to $\zeta_{1/2}(\lambda_1, \lambda_2)$. The bottom row of Figure 6.1 shows that the variance of $\mathbf{z}(\lambda_1, \lambda_2)$ is stable and close to one for a wide range of (λ_1, λ_2). To summarise, the above experiment shows that $\zeta_{1/2}(X_1, X_2)$, the Fisz transform of X_1 and X_2 with exponent $1/2$, can be thought of as an approximately Gaussian variable with mean $\zeta_{1/2}(\lambda_1, \lambda_2)$ and variance bounded above by (and close to) one.

The above discussion concentrates on the properties of individual Fisz-transformed Poisson variables. However, as we observed earlier, the Fisz transform with exponent $1/2$ can be viewed as the division of a Haar detail coefficient by the square root of the corresponding smooth coefficient. Motivated by this observation, we now introduce a full Haar-Fisz transform, where we perform this operation on all Haar detail coefficients of a given vector of Poisson counts.

6.2 The Haar-Fisz transform for Poisson counts

In this section, we provide details of the Haar-Fisz transform, which stabilises the variance of sequences of Poisson counts and brings their distribution closer to normality. The input to the algorithm is a vector $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ for $N = 2^J$, where $v_i \geq 0$ for all i. Typically, $\mathbf{v}$ will be a vector of Poisson counts. The

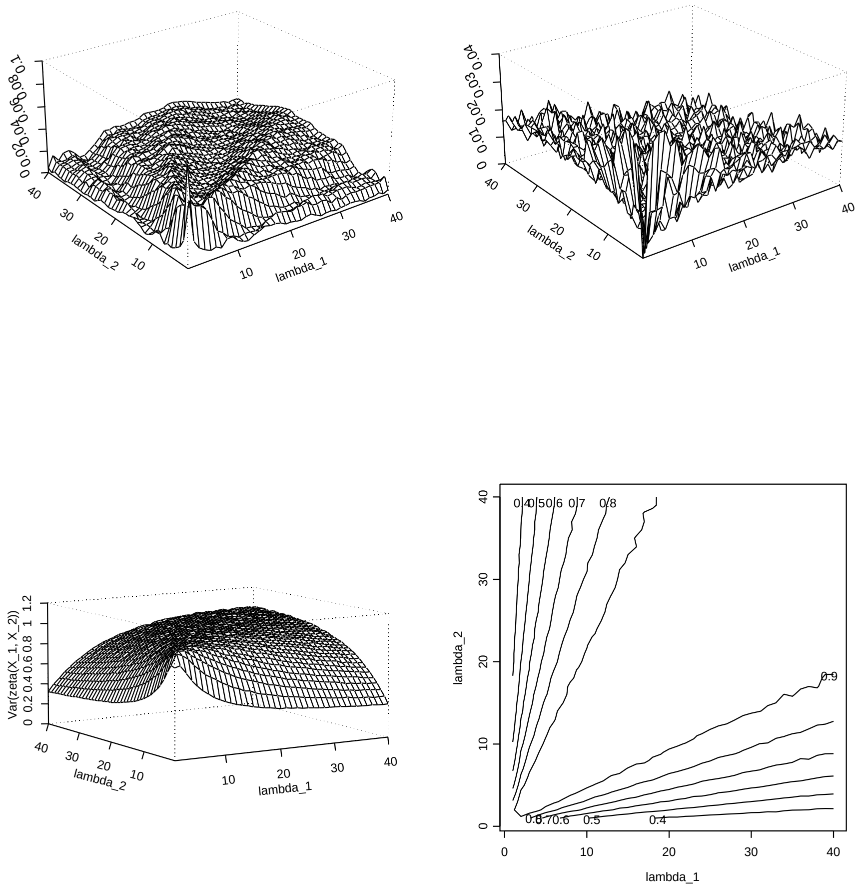

Figure 6.1: Top left: Difference between Kolmogorov-Smirnov test statistics computed on Anscombe-transformed Poisson variables with intensity $\max(\lambda_1, \lambda_2)$, and $\mathbf{z}(\lambda_1, \lambda_2)$. Positive difference means that Haar-Fisz is closer to Gaussian. Top right: $|\bar{\mathbf{z}}(\lambda_1, \lambda_2) - \zeta_{1/2}(\lambda_1, \lambda_2)|$. Bottom left (and right): perspective (and contour) plot of $\mathrm{Var}(\mathbf{z}(\lambda_1, \lambda_2))$.

output of the Haar-Fisz transform is a vector $\mathbf{u} = (u_0, u_1, \ldots, u_{N-1})$, constructed as follows:

1. Let
$$s_n^J = v_n \tag{6.2}$$

2. For each $j = J-1, J-2, \ldots, 0$, recursively form vectors $\mathbf{s}^j$ and $\mathbf{f}^j$:
$$\begin{aligned} s_n^j &= \frac{1}{2}(s_{2n}^{j+1} + s_{2n+1}^{j+1}) \qquad (6.3)\\ f_n^j &= \frac{s_{2n}^{j+1} - s_{2n+1}^{j+1}}{2\sqrt{s_n^j}}, \qquad (6.4)\end{aligned}$$
for $n = 0, 1, \ldots, 2^j - 1$ (with the convention $0/0 = 0$).

3. For each $j = 0, 1, \ldots, J-1$, recursively modify the vectors $\mathbf{s}^{j+1}$:
$$\begin{aligned} s_{2n}^{j+1} &= s_n^j + f_n^j \qquad (6.5)\\ s_{2n+1}^{j+1} &= s_n^j - f_n^j, \qquad (6.6)\end{aligned}$$
for $n = 0, 1, \ldots, 2^j - 1$.

4. Set $\mathbf{u} := \mathbf{s}^J$.

For the purpose of this chapter, denote $\mathcal{F}\mathbf{v} := \mathbf{u}$. The nonlinear operator $\mathcal{F}$ is called the *Haar-Fisz transform of* $\mathbf{v}$. A few important remarks are in order.

- The algorithm is invertible, i.e. $\mathbf{v}$ can be reconstructed from $\mathcal{F}\mathbf{v}$ by reversing the steps 4.–1.

- The steps 2.–4. of the algorithm are similar to the forward and inverse Discrete Haar Transform except the division by $(s_n^j)^{1/2}$ in formula (6.4).

- Formula (6.4) can be written as
$$f_n^j = 2^{-1/2}\zeta_{1/2}(s_{2n}^{j+1}, s_{2n+1}^{j+1}). \tag{6.7}$$
In other words, f_n^j is the (scaled) result of the Fisz transform with exponent 1/2 of two neighbouring smooth coefficients s_{2n}^{m+1} and s_{2n+1}^{m+1}.

- Like the Haar DWT and the Haar-Fisz transform for the wavelet periodogram, the Haar-Fisz transform for Poisson counts is of computational order $O(N)$.

6.2.1 Example

As an example, we demonstrate the Haar-Fisz transform applied to the input vector $\mathbf{v}$ of length 8 now with all positive entries. The Haar-Fisz transform $\mathbf{u} = \mathcal{F}\mathbf{v}$ is given by

$$
\begin{aligned}
u_0 &= \frac{\sum_{i=0}^{7} v_i}{8} + \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} + \frac{v_0 + v_1 - (v_2 + v_3)}{2\sqrt{\sum_{i=0}^{3} v_i}} + \frac{v_0 - v_1}{\sqrt{2}\sqrt{v_0 + v_1}}, \\
u_1 &= \frac{\sum_{i=0}^{7} v_i}{8} + \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} + \frac{v_0 + v_1 - (v_2 + v_3)}{2\sqrt{\sum_{i=0}^{3} v_i}} - \frac{v_0 - v_1}{\sqrt{2}\sqrt{v_0 + v_1}}, \\
u_2 &= \frac{\sum_{i=0}^{7} v_i}{8} + \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} - \frac{v_0 + v_1 - (v_2 + v_3)}{2\sqrt{\sum_{i=0}^{3} v_i}} + \frac{v_2 - v_3}{\sqrt{2}\sqrt{v_2 + v_3}}, \\
u_3 &= \frac{\sum_{i=0}^{7} v_i}{8} + \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} - \frac{v_0 + v_1 - (v_2 + v_3)}{2\sqrt{\sum_{i=0}^{3} v_i}} - \frac{v_2 - v_3}{\sqrt{2}\sqrt{v_2 + v_3}}, \\
u_4 &= \frac{\sum_{i=0}^{7} v_i}{8} - \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} + \frac{v_4 + v_5 - (v_6 + v_7)}{2\sqrt{\sum_{i=4}^{7} v_i}} + \frac{v_4 - v_5}{\sqrt{2}\sqrt{v_4 + v_5}}, \\
u_5 &= \frac{\sum_{i=0}^{7} v_i}{8} - \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} + \frac{v_4 + v_5 - (v_6 + v_7)}{2\sqrt{\sum_{i=4}^{7} v_i}} - \frac{v_4 - v_5}{\sqrt{2}\sqrt{v_4 + v_5}}, \\
u_6 &= \frac{\sum_{i=0}^{7} v_i}{8} - \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} - \frac{v_4 + v_5 - (v_6 + v_7)}{2\sqrt{\sum_{i=4}^{7} v_i}} + \frac{v_6 - v_7}{\sqrt{2}\sqrt{v_6 + v_7}}, \\
u_7 &= \frac{\sum_{i=0}^{7} v_i}{8} - \frac{\sum_{i=0}^{3} v_i - \sum_{i=4}^{7} v_i}{2\sqrt{2}\sqrt{\sum_{i=0}^{7} v_i}} - \frac{v_4 + v_5 - (v_6 + v_7)}{2\sqrt{\sum_{i=4}^{7} v_i}} - \frac{v_6 - v_7}{\sqrt{2}\sqrt{v_6 + v_7}}
\end{aligned} \tag{6.8}
$$

Formulas (6.8) are a special case of the general formula for $\mathcal{F}$ given in the next section.

6.2.2 A general formula for the Haar-Fisz transform

We will now introduce an explicit general formula for the operator $\mathcal{F}$ which is used in the proofs later in this chapter. Let $\mathbf{v} = (v_0, v_1, \dots, v_{N-1})$ be the vector of Poisson counts, and let $\mathbf{u} = (u_0, u_1, \ldots, u_{N-1})$ be the Haar-Fisz transform of $\mathbf{v}$: $\mathbf{u} = \mathcal{F}\mathbf{v}$. Bearing in mind that N is an integer power of two, we denote $J = \log_2(N)$. We introduce the family of Haar wavelet vectors $\{\psi^{j,k}\}$, where $j = 0, 1, \ldots, J-1$ is the scale parameter, and $k = l2^{J-j}, l = 0, 1, \ldots, j$, is the location parameter. The components of $\psi^{j,k}$ will be denoted by $\psi_n^{j,k}$, for $n = 0, 1, \ldots, N-1$. We define

$$\psi_n^{j,k} = \begin{cases} 0 & \text{for} \quad n < k \\ 1 & \text{for} \quad k \le n < k + 2^{J-j-1} \\ -1 & \text{for} \quad k + 2^{J-j-1} \le n < k + 2^{J-j} \\ 0 & \text{for} \quad k + 2^{J-j} \le n. \end{cases} \tag{6.9}$$

Similarly, we introduce the family of Haar scaling vectors $\{\phi^{j,k}\}$, whose components will be denoted by $\phi_n^{j,k}$ (the range of j, k, and n remains unchanged). We define

$$\phi_n^{j,k} = \begin{cases} 0 & \text{for} \quad n < k \\ 1 & \text{for} \quad k \le n < k + 2^{J-j} \\ 0 & \text{for} \quad k + 2^{J-j} \le n. \end{cases} \tag{6.10}$$

Our definition of discrete Haar wavelets is similar to that of Nason *et al.* (2000), Section 2. The difference is that we "pad" the wavelet vectors with zeros on both sides so that they all have length N, and we do not normalise them.

Further, let $\langle \cdot, \cdot \rangle$ denote the inner product of two vectors, and let $\mathbf{b}^J(n) = (b_0^J(n), b_1^J(n), \ldots, b_{J-1}^J(n))$ be the binary representation of the integer n, where $n < 2^J$.

The formula for the nth element of $\mathbf{u} = \mathcal{F}\mathbf{v}$ is

$$u_n = \frac{\langle \phi^{0,0}, \mathbf{v} \rangle}{N} + \sum_{j=0}^{J-1} (-1)^{b_j^J(n)} 2^{\frac{j-J}{2}} c_{j,J,n}(\mathbf{v}), \tag{6.11}$$

where

$$c_{j,J,n} = \begin{cases} \frac{\langle \psi^{j, \lfloor n/2^{J-j} \rfloor 2^{J-j}}, \mathbf{v} \rangle}{\langle \phi^{j, \lfloor n/2^{J-j} \rfloor 2^{J-j}}, \mathbf{v} \rangle^{\frac{1}{2}}} & \text{if} \quad \langle \phi^{j, \lfloor n/2^{J-j} \rfloor 2^{J-j}}, \mathbf{v} \rangle > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.12}$$

6.3 Properties of the Haar-Fisz transform for constant intensities

In this section, we state and prove two propositions concerning the asymptotic behaviour of $\mathcal{F}\mathbf{v}$, where $\mathbf{v}$ is a Poisson vector of constant intensity. The case of non-constant intensities will be considered in Section 6.4. Proposition 6.3.1 says that the coefficients of $\mathcal{F}\mathbf{v}$ are asymptotically uncorrelated, and Proposition 6.3.2 says that they are also asymptotically normal with variance one.

Proposition 6.3.1 *Let $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ be a vector of i.i.d. Poisson variables with mean λ, and let N be an integer power of two. Let $\mathbf{u} = \mathcal{F}\mathbf{v}$. For $m \neq n$, we have*

$$cor(u_m, u_n) \to 0 \quad as \quad \lambda \to \infty \quad and \quad \lambda/N \to 0. \tag{6.13}$$

Proof. We will first calculate the correlation between the modified detail coefficients at two different scales. The detail coefficient at any given scale has the form

$$D_f = (X_0 - X_1) f(X_0 + X_1)$$

where X_0 and X_1 are some independent, identically distributed Poisson variables, and $f(x) = x^{-1/2}$ with $f(0) = 0$. The detail coefficient at any coarser scale depends on X_0, X_1 *through their sum only*, i.e. we have

$$D_c = g(X_0 + X_1),$$

where g also depends on some other Poisson variables X_i, $i \neq 0, 1$. Since X_0, X_1 are identically distributed, we obviously have

$$\mathbb{E}(D_f) = 0 \text{ and } \mathbb{E}(D_f D_c) = 0, \tag{6.14}$$

and so $\mathrm{cov}(D_f, D_c) = 0$. We can show in a similar way that the smooth coefficient $\langle \phi^{0,0}, \mathbf{v} \rangle / N$ is uncorrelated with any of the detail coefficients.

We are now in a position to calculate $\mathrm{cov}(u_m, u_n)$. From formula (6.11) it is clear that the variables will share the "smooth" term $\langle \phi^{0,0}, \mathbf{v} \rangle / N$, which we will denote by μ to simplify the notation. Since the integer $\lfloor n/2^{J-j} \rfloor 2^{J-j}$ (see formula (6.11)) depends only on the first j bits in the binary expansion of n, the variables u_m and u_n will also share the term

$$X := \sum_{j=0}^{J^*-1} (-1)^{b_j^J(n)} 2^{\frac{j-J}{2}} c_{j,J,n}(\mathbf{v}), \tag{6.15}$$

where $J^* = \min\{j : b_j^J(n) \neq b_j^J(m)\}$. Using the definition in formula (6.12), it can be proved that

$$(-1)^{b_{J^*}^J(m)} 2^{\frac{J^*-J}{2}} c_{J^*,J,m}(\mathbf{v}) = -(-1)^{b_{J^*}^J(n)} 2^{\frac{J^*-J}{2}} c_{J^*,J,n}(\mathbf{v}). \tag{6.16}$$

The term on the LHS of equation (6.16) will be denoted by Y. We also denote

$$\begin{aligned} Z_1 &= \sum_{j=J^*+1}^{J-1} (-1)^{b_j^J(m)} 2^{\frac{j-J}{2}} c_{j,J,m}(\mathbf{v}) \\ Z_2 &= \sum_{j=J^*+1}^{J-1} (-1)^{b_j^J(n)} 2^{\frac{j-J}{2}} c_{j,J,n}(\mathbf{v}). \end{aligned}$$

It takes a closer look at formula (6.11) to see that Z_1 and Z_2 are independent (they are functions of *different* components of $\mathbf{v}$). Using the formulas in (6.14), we now write

$$\begin{aligned} \mathrm{cov}(u_m, u_n) &= \mathrm{cov}(\mu + X - Y + Z_1, \mu + X + Y + Z_2) \\ &= \mathrm{Var}(\mu) + \mathrm{Var}(X) - \mathrm{Var}(Y). \end{aligned}$$

For λ large enough, as X and Y become approximately normal (see Fisz (1955)), we have

$$\begin{aligned} \mathrm{Var}(X) &\leq \sum_{j=0}^{J^*-1} (1+\epsilon) 2^{j-J} = (1+\epsilon)(2^{J^*-J} - 2^{-J}) \\ \mathrm{Var}(Y) &\geq (1-\epsilon) 2^{J^*-J} \end{aligned}$$

Moreover, we have $\mathrm{Var}(\mu) = \lambda/N \to 0$ by assumption. Since $N = 2^J \to \infty$, we have

$$\mathrm{cov}(u_m, u_n) \leq \lambda/N + (1+\epsilon)(2^{J^*-J} - 2^{-J}) - (1-\epsilon)2^{J^*-J} \to 0$$

as $\epsilon \to 0$ (note that 2^{J^*-J} is constant), which completes the proof. □

Proposition 6.3.2 *Let* $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ *be a vector of i.i.d. Poisson variables with mean* λ*, and let* N *be an integer power of two. Let* $\mathbf{u} = \mathcal{F}\mathbf{v}$*. For all* $n = 0, 1, \ldots, N-1$*, we have*

$$u_n - \lambda = \nu + Y_n, \tag{6.17}$$

where

$$\begin{aligned} \nu &\xrightarrow{D} 0 \quad as \quad \lambda/N \to 0 \\ Y_n &\xrightarrow{D} N(0,1) \quad as \quad (\lambda, N) \to (\infty, \infty). \end{aligned} \tag{6.18}$$

Proof. Without loss of generality, let us concentrate on u_0. Let $J = \log_2(N)$, and let us denote

$$W_j(\lambda) = \begin{cases} \frac{\sum_{i=0}^{2^j-1} v_i - \sum_{i=2^j}^{2^{j+1}-1} v_i}{\sqrt{\sum_{i=0}^{2^{j+1}-1} v_i}} & \text{if} \quad \sum_{i=0}^{2^{j+1}-1} v_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.19}$$

to emphasise the dependence of W_j on λ. The following equality holds (see the example in Section 6.2.1, and formulas (6.11) and (6.12))

$$u_0 = N^{-1} \sum_{i=0}^{N-1} v_i + \sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda). \tag{6.20}$$

Set

$$\nu = N^{-1} \sum_{i=0}^{N-1} v_i - \lambda \text{ and } Y_0 = \sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda).$$

We will first show that $Y_0 \xrightarrow{D} N(0,1)$ as $(\lambda, J) \to (\infty, \infty)$. Let us fix $\epsilon_1 > 0$. By Theorem 5.1.1, if λ or j are large enough, then we have

$$\mathrm{Var}(W_j(\lambda)) = (1 + \epsilon_j^\lambda) \le 1 + \epsilon_1, \tag{6.21}$$

where $|\epsilon_j^\lambda| < \epsilon_1$. Also, for all λ, the variables $W_j(\lambda)$ are uncorrelated (see the proof of Proposition 6.3.1).

Using the symmetry of $W_j(\lambda)$, the Chebyshev inequality, the orthogonality of $W_j(\lambda)$, and formula (6.21), for large λ, J, and $M > J$ we have

$$\begin{aligned} P\left(\sum_{j=J}^{M-1} 2^{\frac{-j-1}{2}} W_j(\lambda) < -\epsilon\right) &= P\left(\sum_{j=J}^{M-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > \epsilon\right) \\ &\leq \epsilon^{-2}\,\mathrm{Var}\left(\sum_{j=J}^{M-1} 2^{\frac{-j-1}{2}} W_j(\lambda)\right) \\ &= \epsilon^{-2} \sum_{j=J}^{M-1} 2^{-j-1}\mathrm{Var}\,(W_j(\lambda)) \\ &\leq \epsilon^{-2} \sum_{j=J}^{\infty} 2^{-j-1}\mathrm{Var}\,(W_j(\lambda)) \\ &\leq \epsilon^{-2}(1+\epsilon_1)2^{-J}. \end{aligned}$$

Clearly, we have that

$$\forall\,\epsilon \quad \exists\,J_0 \quad \forall\,J \geq J_0 \quad \epsilon^{-2}(1+\epsilon_1)2^{-J} \leq \epsilon. \tag{6.22}$$

Observe now that

$$\forall\,J \quad \sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) \stackrel{D}{\to} N(0, 1-2^{-J}) \quad \text{as} \quad \lambda \to \infty. \tag{6.23}$$

Here we have a finite linear combination of orthogonal variables, each of which converges in distribution to $N(0,1)$ by Theorem 5.1.1. The finite linear combination will therefore converge to the finite linear combination of orthogonal (= independent) normal variables, whose variances sum up to $1 - 2^{-J}$. Denote by $S_{\sigma^2}(t)$ the survival function of a normal variable with mean zero and variance σ^2. Note two properties of the family $\{S_{1-2^{-J}}(t)\}_{J=1}^{\infty}$: $\|S_{1-2^{-J}}(\cdot) - S_1(\cdot)\|_\infty \to 0$ as $J \to \infty$; $\{S_{1-2^{-J}}(t)\}_{J=1}^{\infty}$ is uniformly Lipschitz continuous with Lipschitz constant $L = 1/\sqrt{\pi}$.

Now fix $\epsilon > 0$ and choose the corresponding J_0 in (6.22). For an arbitrary fixed t, examine the difference

$$D_1 = \left| P\left(\sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t\right) - S_1(t) \right|. \tag{6.24}$$

We have

$$
\begin{aligned}
&P\left(\sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t\right) = \\
&\quad P\left(\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) + \sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t\right) \le \\
&\quad P\left(\left\{\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t-\epsilon\right\} \vee \left\{\sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > \epsilon\right\}\right) \le \\
&\quad P\left(\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t-\epsilon\right) + P\left(\sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > \epsilon\right) \le \\
&\quad (S_{1-2^{-J_0}}(t-\epsilon)+\epsilon)+\epsilon \le S_{1-2^{-J_0}}(t) + \epsilon/\sqrt{\pi} + 2\epsilon \le \\
&\quad S_1(t) + \epsilon + \left(1/\sqrt{\pi}+2\right)\epsilon \le S_1(t) + 4\epsilon.
\end{aligned} \tag{6.25}
$$

On the other hand, we have

$$
\begin{aligned}
&P\left(\sum_{j=0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t\right) \ge \\
&\quad P\left(\left\{\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t+\epsilon\right\} \wedge \left\{\sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > -\epsilon\right\}\right) = \\
&\quad P\left(\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t+\epsilon\right) + P\left(\sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > -\epsilon\right) - \\
&\quad P\left(\left\{\sum_{j=0}^{J_0-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > t+\epsilon\right\} \vee \left\{\sum_{j=J_0}^{J-1} 2^{\frac{-j-1}{2}} W_j(\lambda) > -\epsilon\right\}\right) \ge \\
&\quad (S_{1-2^{-J_0}}(t+\epsilon)-\epsilon) + (1-\epsilon) - 1 \ge S_{1-2^{-J_0}}(t) - \frac{1}{\sqrt{\pi}}\epsilon - 2\epsilon \ge \\
&\quad S_1(t) - \epsilon - \left(\frac{1}{\sqrt{\pi}}+2\right)\epsilon \ge S_1(t) - 4\epsilon.
\end{aligned} \tag{6.26}
$$

Inequalities (6.25) and (6.26) together prove that the difference D_1 of formula (6.24) is arbitrarily small for λ and J large enough, which proves the convergence.

We will now show that $\nu \xrightarrow{D} 0$ as $\lambda/N \to 0$. We denote by $S^0(t)$ the survival function of the constant variable 0. Consider the difference

$$
D_2 = \left| P\left(N^{-1}\sum_{i=0}^{N-1} v_i - \lambda > t\right) - S^0(t)\right|. \tag{6.27}
$$

For $t > 0$, we have

$$\begin{aligned} D_2 &= P\left(N^{-1}\sum_{i=0}^{N-1}(v_i-\lambda) > t\right) \le N^{-2}t^{-2}\,\mathbb{E}\left(\sum_{i=0}^{N-1}(v_i-\lambda)\right)^2 \\ &= N^{-2}t^{-2}\sum_{i=0}^{N-1}\mathrm{Var}(v_i) = N^{-1}t^{-2}\lambda \to 0 \quad \text{as} \quad \lambda/N \to 0. \end{aligned} \tag{6.28}$$

For $t < 0$, we have

$$\begin{aligned} D_2 &= \left|P\left(N^{-1}\sum_{i=0}^{N-1} v_i - \lambda > -|t|\right) - 1\right| = P\left(N^{-1}\sum_{i=0}^{N-1} v_i - \lambda \le -|t|\right) \\ &= P\left(-N^{-1}\sum_{i=0}^{N-1} v_i + \lambda \ge |t|\right) \le N^{-1}t^{-2}\lambda \to 0 \quad \text{as} \quad \lambda/N \to 0. \end{aligned} \tag{6.29}$$

Inequalities (6.28) and (6.29) show that $\nu \xrightarrow{D} 0$ as $\lambda/N \to 0$. The proof of Proposition 6.3.2 is completed. □

6.4 Properties of the Haar-Fisz transform for non-constant intensities

In this section, we assess the degree of Gaussianisation and variance stabilisation provided by the Haar-Fisz transform for non-constant Poisson intensities. Also, we examine the amount of correlation between the Haar-Fisz transformed variables.

6.4.1 Decorrelation and Gaussianisation

We begin by empirically investigating the degree of correlation between the Haar-Fisz transformed variables, as well as their proximity to normality. The details of the computational experiment are as follows: we selected 4 "templates" (vectors of length 128), and shifted each of them by 1/10, 1, 2, 3, and 4 to create $4 \times 5 = 20$ test intensity vectors. The templates are plotted in Figure 6.2.

The template `v0` is used to create constant intensities of 1/10, 1, 2, 3, and 4. Each of the templates `v25`, `v50` and `v75`, after shifting upwards by c (where $c \in \{1/10, 1, 2, 3, 4\}$), becomes a non-constant intensity vector in the shape of a symmetric rectangular hat whose middle part is elevated to the level of $8 + c$ (so

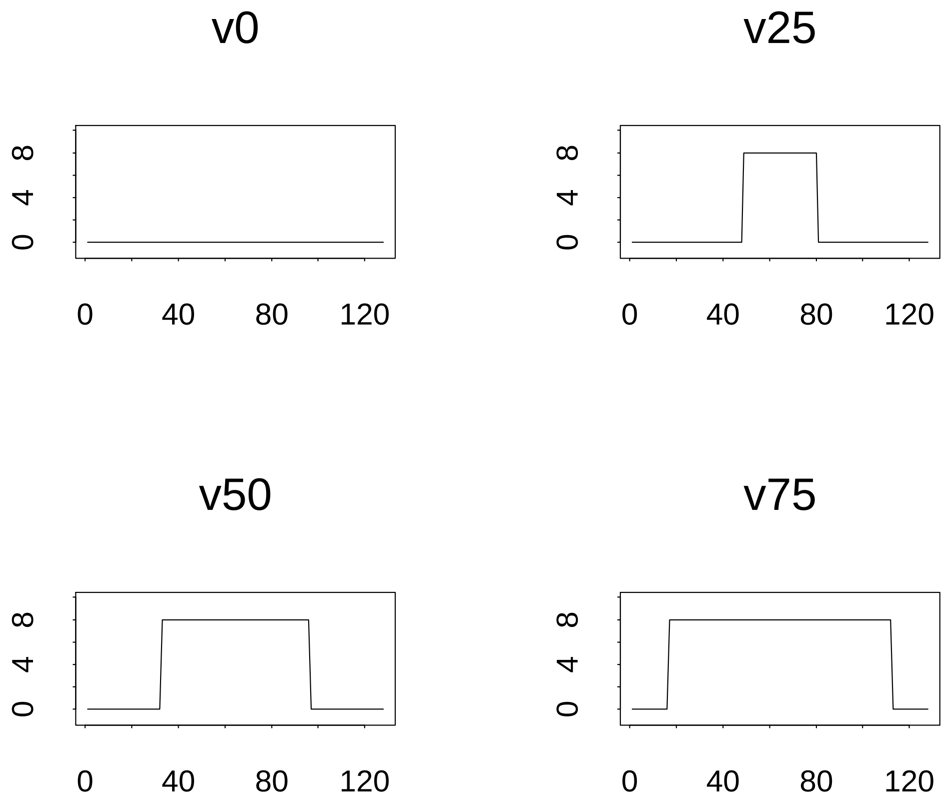

Figure 6.2: Templates used in the experiment of Sections 6.4.1 and 6.4.2.

that it corresponds to a "high" intensity), with the outer parts remaining at the minimum level of c (so that they correspond to a "low" intensity). In the template vn, the elevated middle part stretches over $n\%$ of the length of the template.

Results for the templates `v0`, `v25`, `v50` and `v75` are plotted in Figures 6.3, 6.4, 6.5 and 6.6, respectively. In each of the figures, the consecutive rows correspond to shifts of the corresponding template by 1/10, 1, 2, 3 and 4, respectively.

The left subfigure in each row shows the quantiles of the distribution of $\mathcal{F}\mathbf{x} - \mathcal{F}\lambda$ against the standard normal, where λ is the respective intensity vector and $\mathbf{x}$ is a sample path simulated from it. The quantiles have been averaged over 100 simulated sample paths, rather than basing the result on only one simulation. The right subfigure in each row shows the autocorrelation function of $\mathcal{F}\mathbf{x} - \mathcal{F}\lambda$ for lags 1 to 21, again averaged over 100 simulated sample paths. The corresponding 95% confidence bands have been adjusted to take into account the averaging.

Decorrelation. The decorrelation appears to work fine except for the most non-Gaussian setups: `v0`+1/10, `v25`+1/10 and `v50`+1/10. Also, a tendency can be observed for the sample autocorrelation to be negative rather than positive, even if it is non-significant.

Gaussianisation. Let $\underline{\lambda}$ denote the minimum of the given intensity vector. The experiment demonstrates that while for constant intensities, the degree of Gaussianisation is only satisfactory from about $\underline{\lambda} = 4$ upwards, the "acceptable" level of $\underline{\lambda}$ falls as low as 2 or even less in the case of `v75`. This is so because the stretch of high intensity "makes up" for the failure of the asymptotic mechanism for the period of low intensity.

The following example compares the Gaussianisation property of the Haar-Fisz transform, $\mathcal{F}$, Anscombe's transform, $\mathcal{A}$, and the identity transform. Let us consider the intensity as in the top plot of Figure 6.7 (a rescaled and shifted version of the Donoho & Johnstone (1994) bumps function). This intensity vector will be denoted by λ, and $\mathbf{v}$ will denote a sample path generated from it.

Figure 6.7 compares the Q-Q plots of $\mathbf{v}-\lambda$, $\mathcal{A}\mathbf{v}-\mathcal{A}\lambda$ and $\mathcal{F}\mathbf{v}-\mathcal{F}\lambda$ averaged over 100 samples of $\mathbf{v}$. Clearly, the Q-Q plot shows that the Haar-Fisz transformation does a better job in Gaussianisation. In particular, the Haar-Fisz transformed data is less "stepped" and looks more like variates from a continuous distribution than a discrete one. The Anscombe-transformed data appears more "stepped" for lower quantiles than for higher ones. Further, the tails for Haar-Fisz are more normal than for Anscombe which in turn is more normal than the raw count data.

6.4.2 Variance stabilisation

In this section, we use the templates `v0`, `v25`, `v50` and `v75` to investigate the variance-stabilising properties of the Haar-Fisz transform and the Anscombe transform.

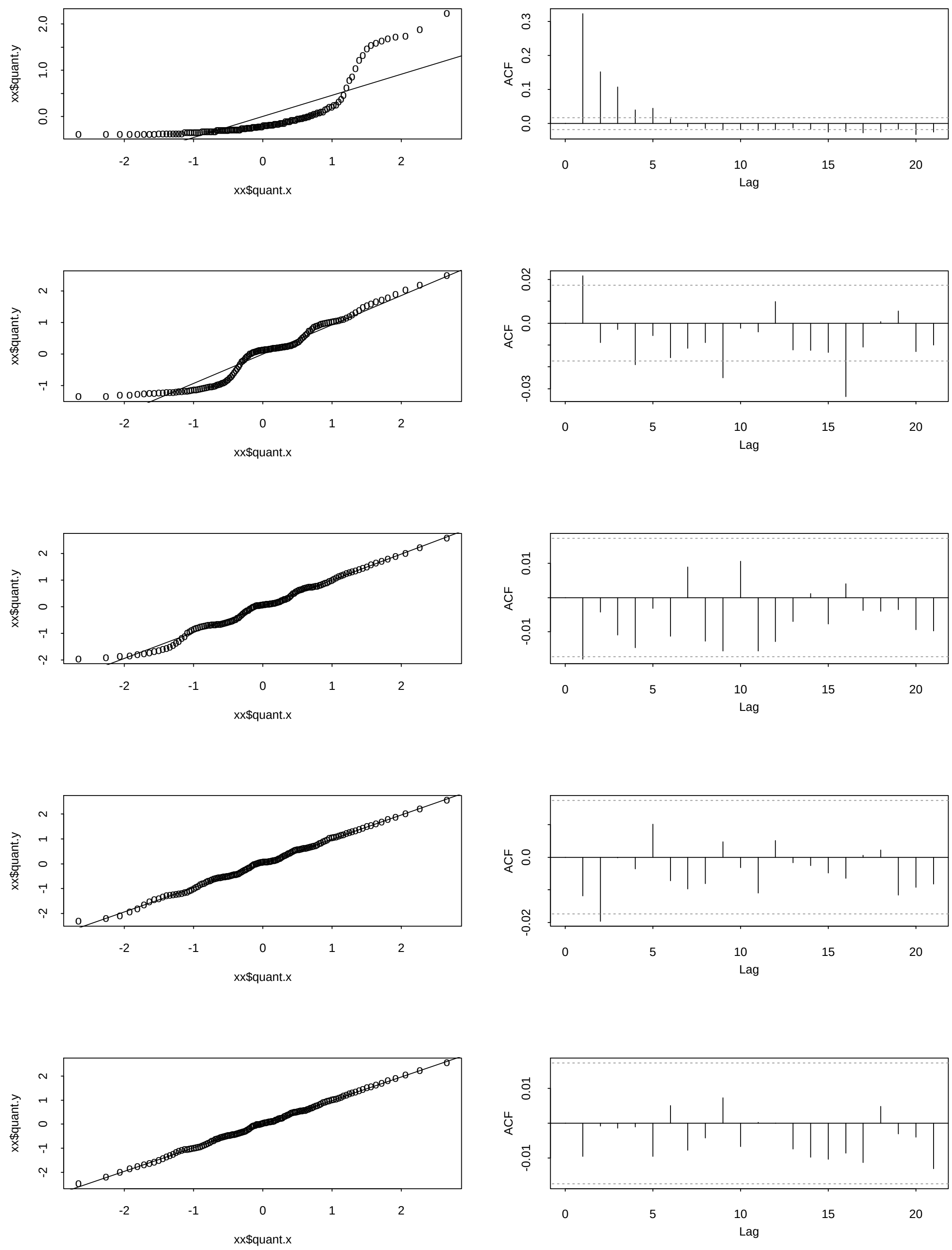

Figure 6.3: Q-Q and acf plots for v0; see Section 6.4.1 for detailed description.

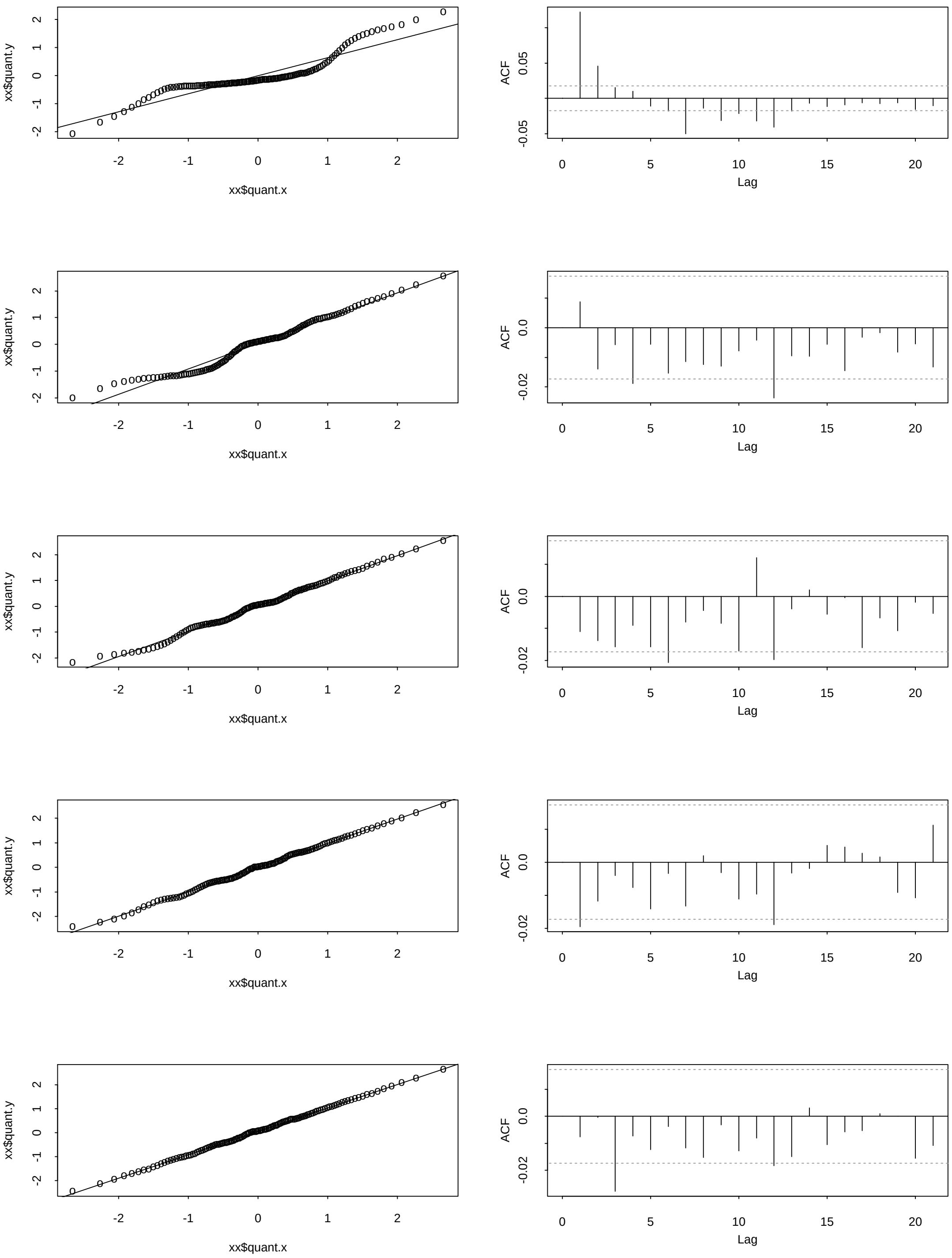

Figure 6.4: Q-Q and acf plots for `v25`; see Section 6.4.1 for detailed description.

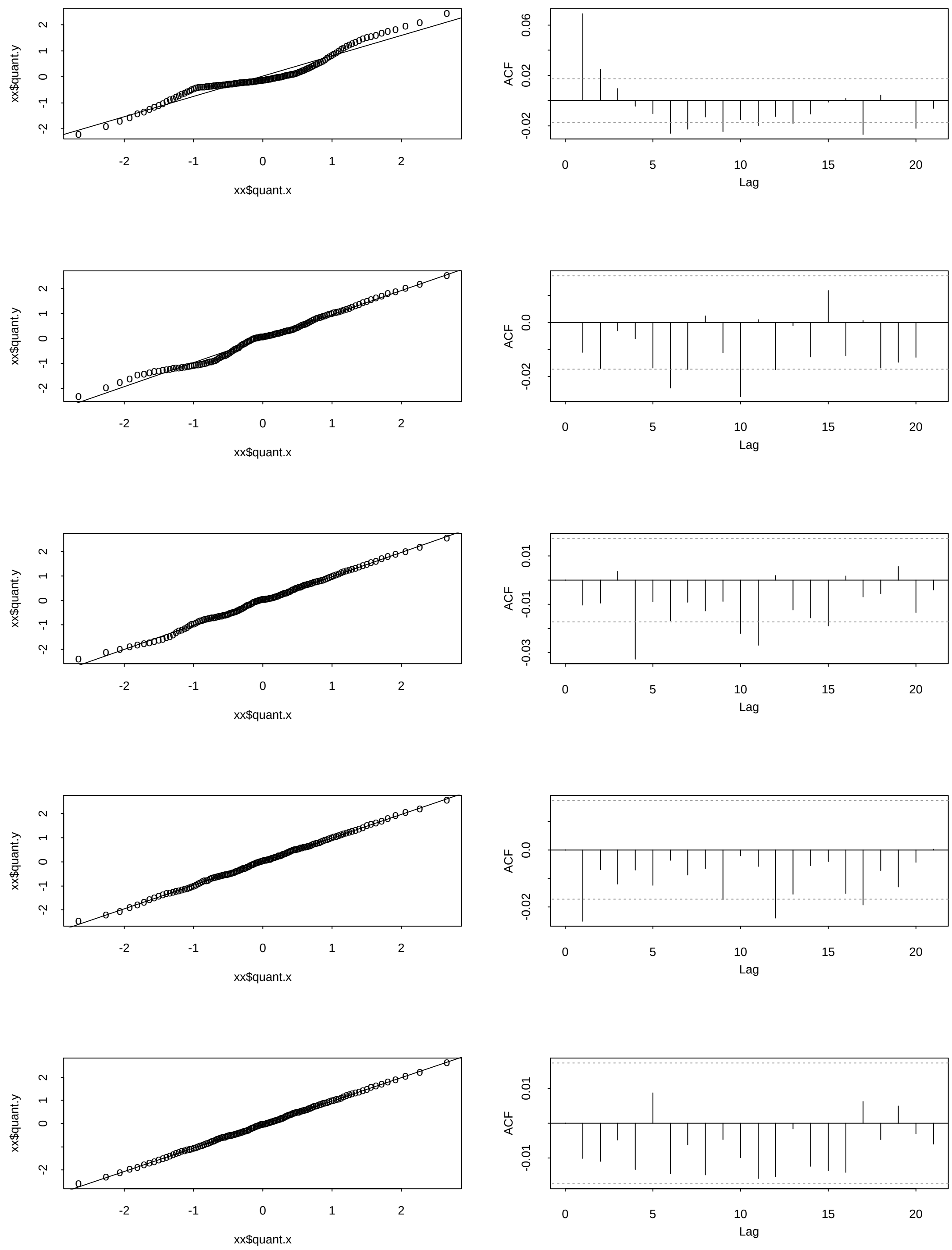

Figure 6.5: Q-Q and acf plots for `v50`; see Section 6.4.1 for detailed description.

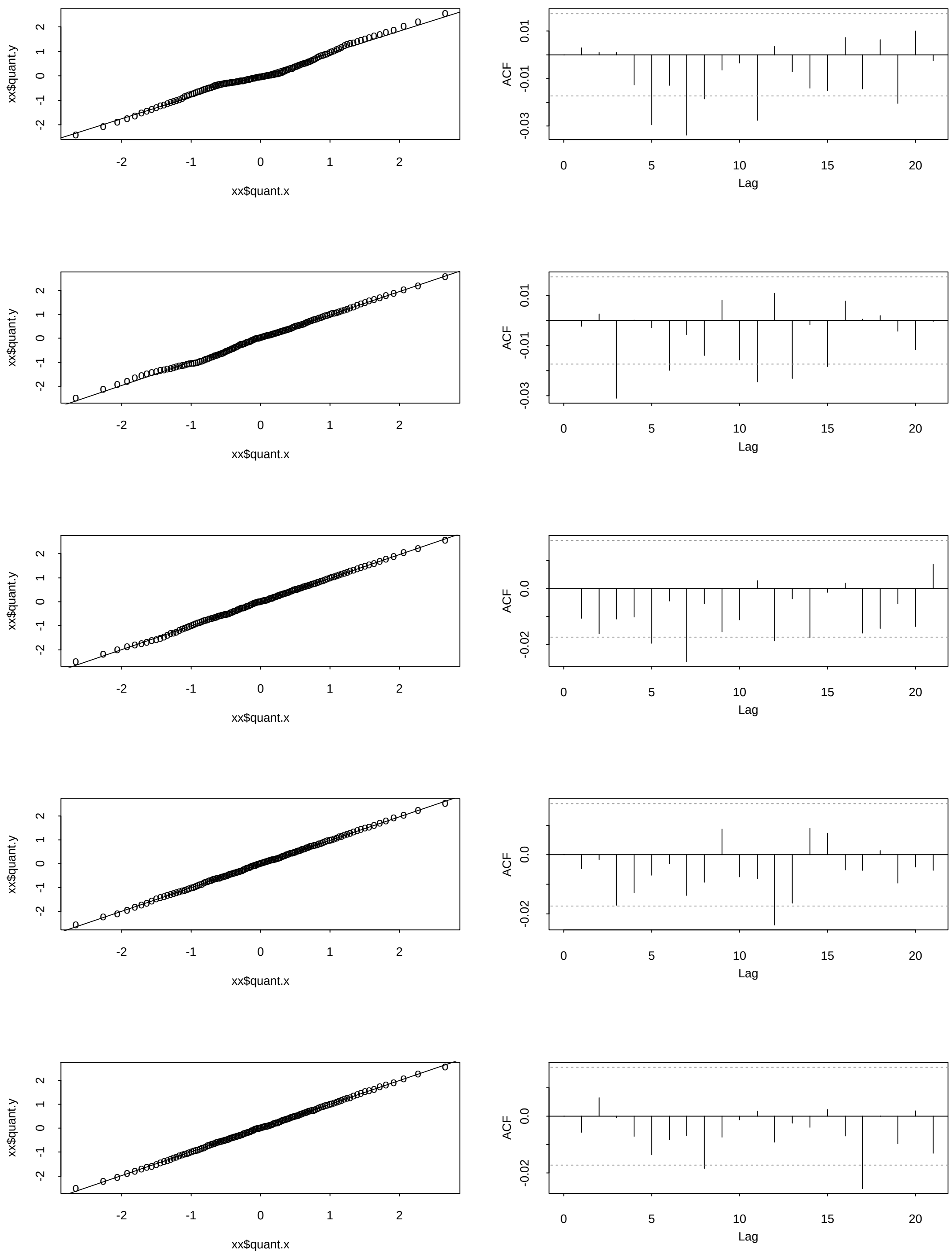

Figure 6.6: Q-Q and acf plots for v75; see Section 6.4.1 for detailed description.

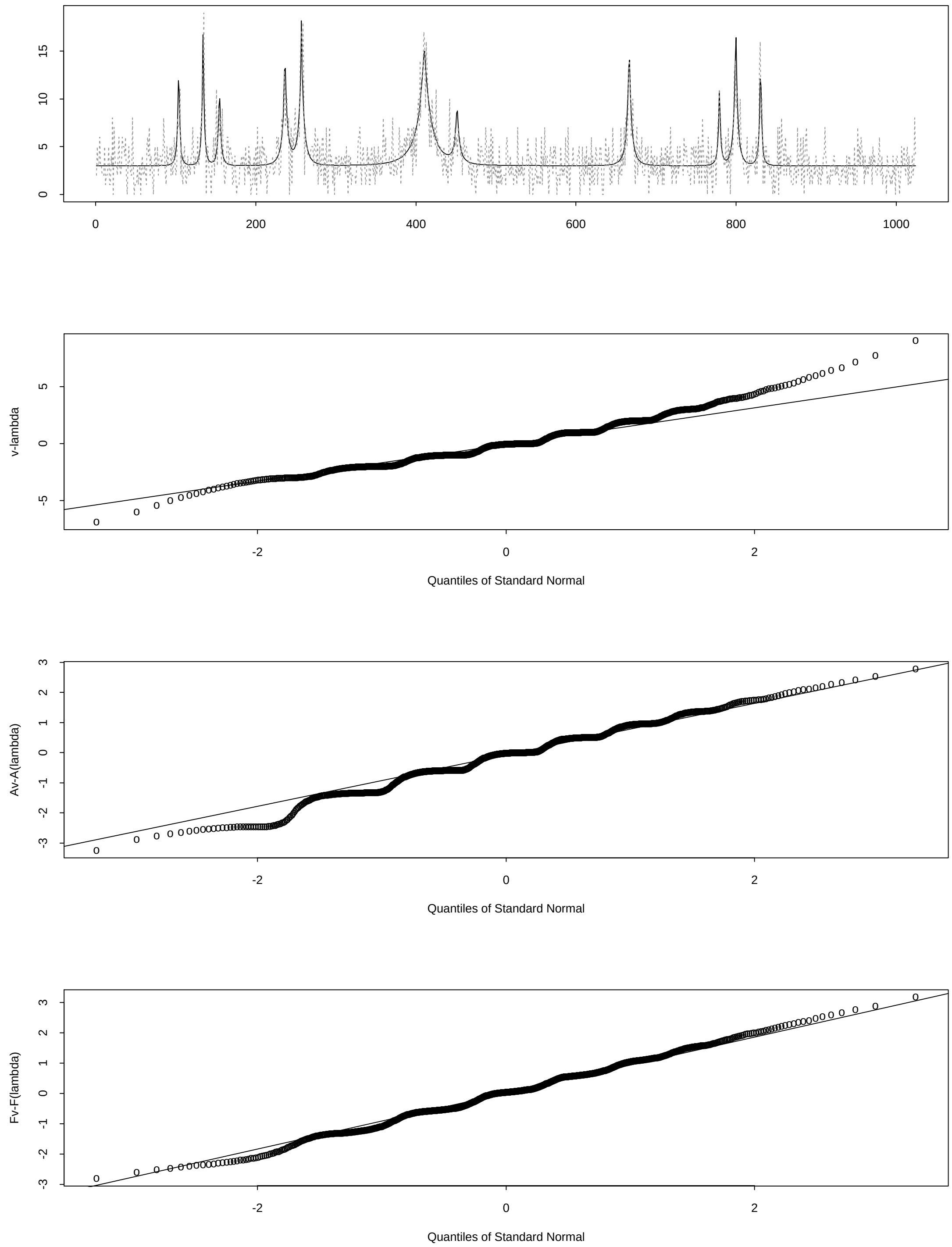

Figure 6.7: From top to bottom: intensity vector λ of Donoho & Johnstone (1994) bumps function (solid; shifted and scaled so that the minimum intensity is 3 and the maximum is 18) and one sample path $\mathbf{v}$ (dotted); Q-Q plots of vectors $\mathbf{v} - \lambda$, $\mathcal{A}\mathbf{v} - \mathcal{A}\lambda$, and $\mathcal{F}\mathbf{v} - \mathcal{F}\lambda$ averaged over 100 $\mathbf{v}$ samples.

Results for the templates v0, v25, v50 and v75 are plotted in Figures 6.8, 6.9, 6.10 and 6.11, respectively. Like in the previous section, the consecutive rows in each of the figures correspond to shifts of the corresponding template by 1/10, 1, 2, 3 and 4, respectively.

The left subfigure in each row shows the squared residuals $(\mathcal{A}\lambda - \mathcal{A}\mathbf{x})^2$, where λ is the respective intensity vector and $\mathbf{x}$ is a sample path simulated from it. The results have been averaged over 1000 simulated sample paths to give an idea of the variance of the Anscombe-transformed vector for each intensity function. The right subfigure shows the analogous quantity for the Haar-Fisz operator $\mathcal{F}$.

The simulation shows that in most cases, Haar-Fisz provides a better variance stabilisation than Anscombe, i.e. the squared residuals are closer to one. For the Haar-Fisz transform, the degree of variance stabilisation seems to be satisfactory from $\underline{\lambda} = 1$ onwards, while for the Anscombe transform — only from about $\underline{\lambda} = 2$. Moreover, for the Haar-Fisz transform, the level of the squared residuals almost exactly reflects the shape of the underlying intensity function (a phenomenon which is natural for the Anscombe transform, but not necessarily expected of Haar-Fisz).

6.4.3 Summary of conclusions

An analogous simulation study with the four templates was also carried out for $N = 1024$. Let "D-G-S" denote the decorrelation, Gaussianisation and variance stabilisation properties of the Haar-Fisz transform. The main conclusions from the two experiments can be summarised as follows.

1. The degree of D-G-S was strikingly similar for sample sizes of $N = 128$ and $N = 1024$: we suspect that the degree of D-G-S is not strongly dependent on N. We consider $N = 128$ to be a short vector in this situation.

2. The greater the minimum of the intensity vector, λ, the higher the degree of D-G-S. For constant intensities D-G-S is *extremely* effective from about $\lambda = 4$.

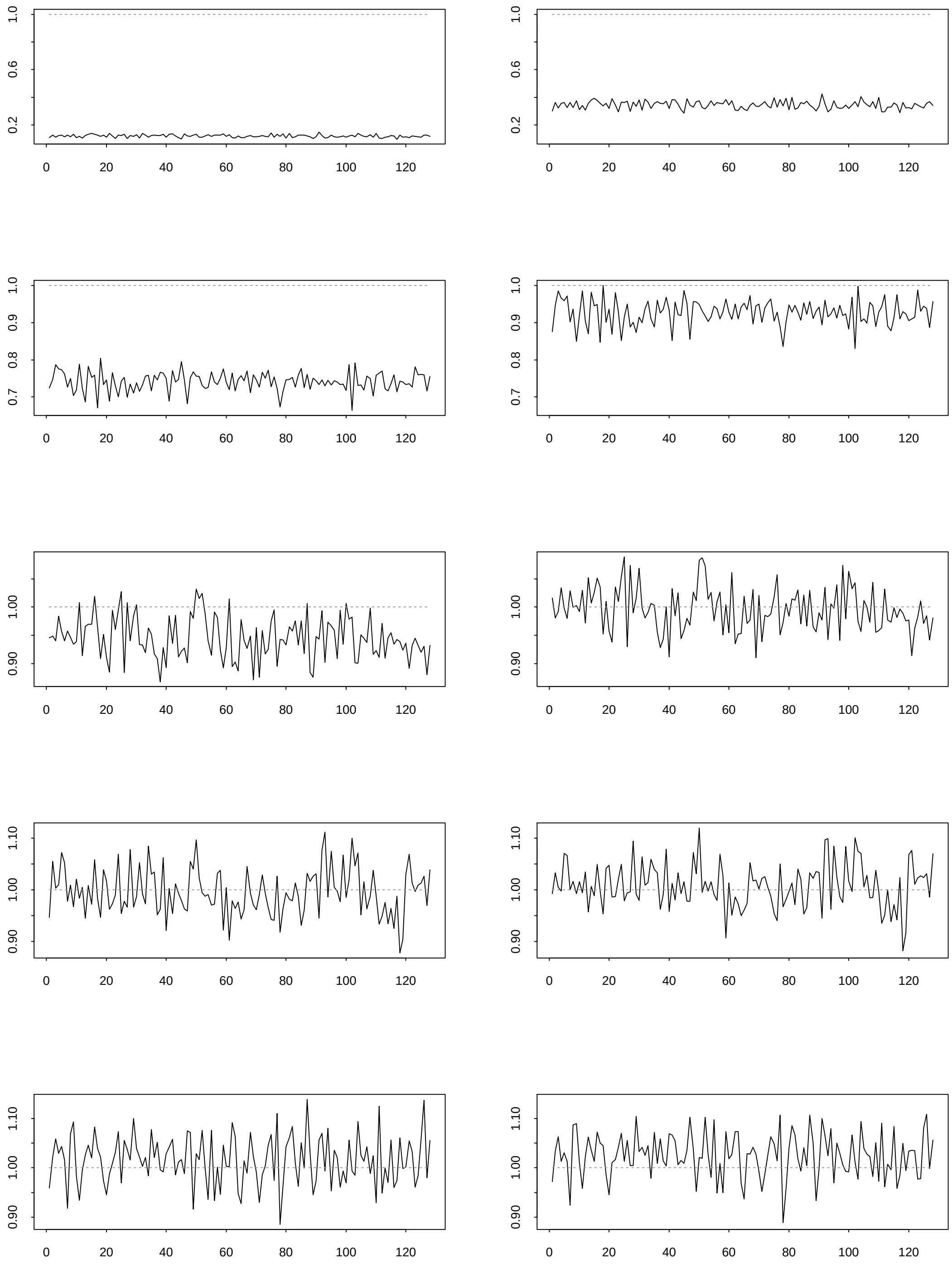

Figure 6.8: Averaged squared residuals for v0; see Section 6.4.2 for detailed description.

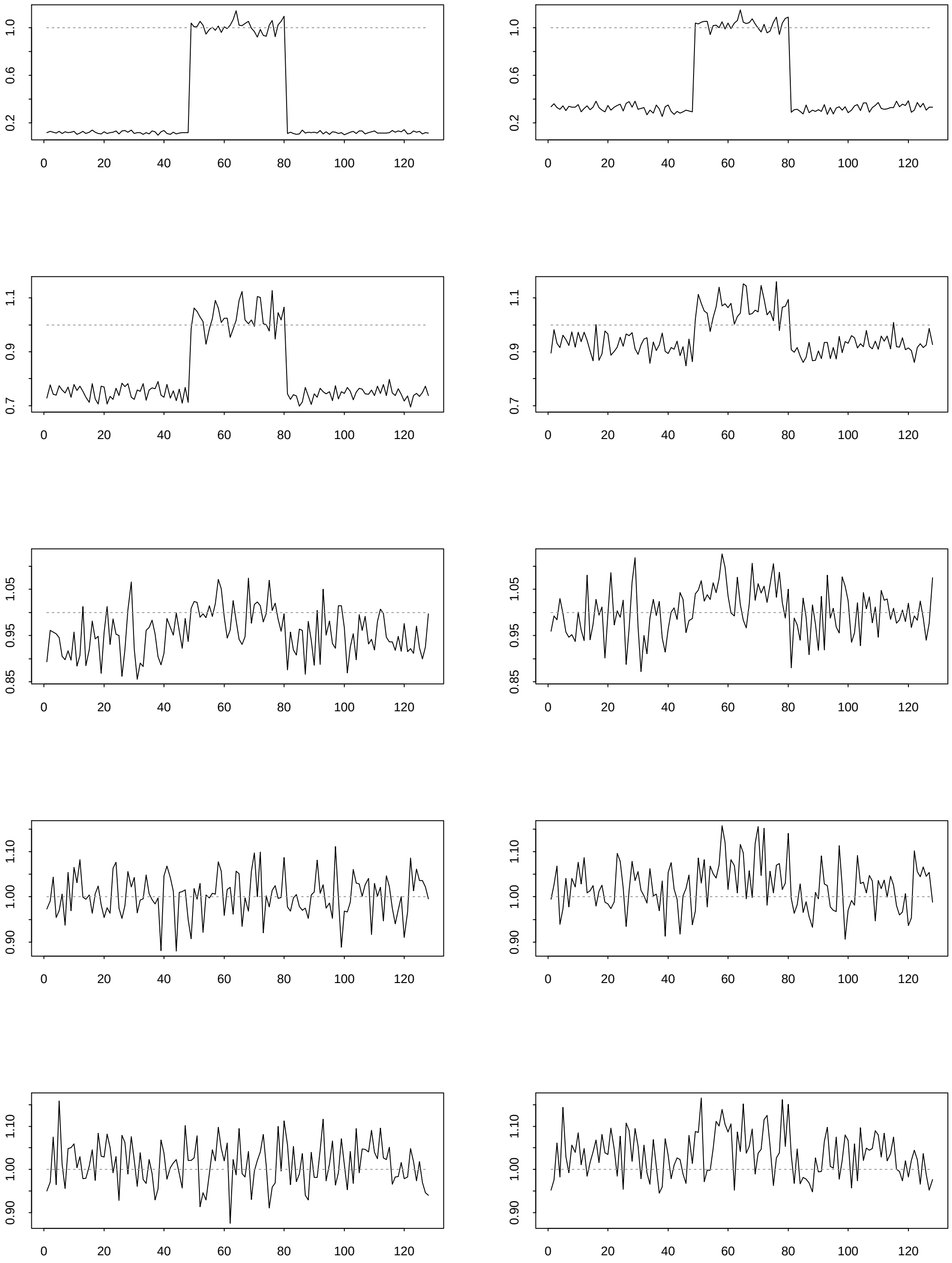

Figure 6.9: Averaged squared residuals for `v25`; see Section 6.4.2 for detailed description.

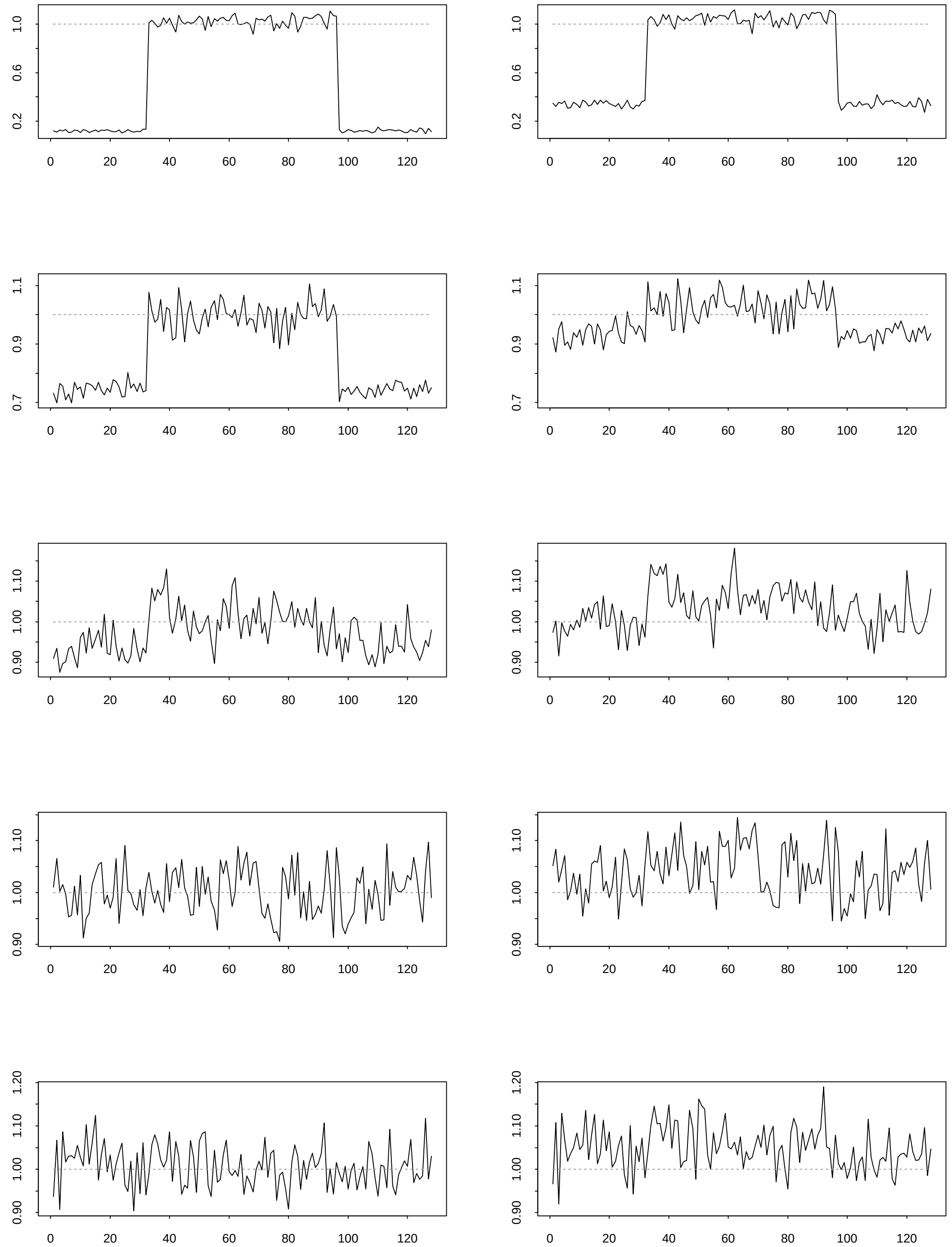

Figure 6.10: Averaged squared residuals for `v50`; see Section 6.4.2 for detailed description.

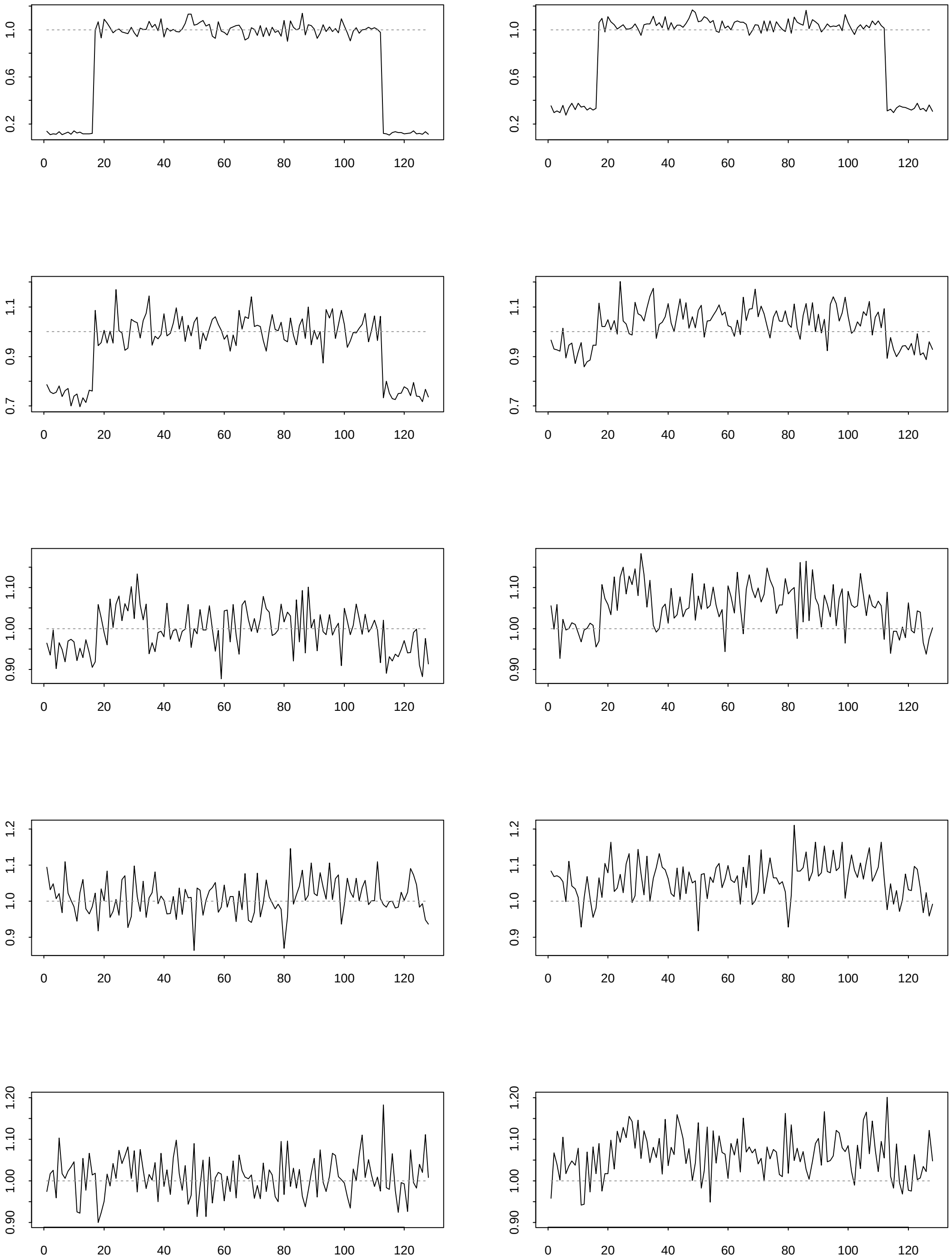

Figure 6.11: Averaged squared residuals for `v75`; see Section 6.4.2 for detailed description.

3. For non-constant intensities, the degree of D-G-S depends not only on $\underline{\lambda} := \min \lambda$ but also on the length of the stretch where the intensity is equal, or close, to $\underline{\lambda}$. The shorter the stretch, the lower the "acceptable" value of $\underline{\lambda}$ for which D-G-S is still very effective. For example, if the intensity is at its constant minimum, 2, for 25% of the time and the remaining intensity is constant at 10, then the D-G-S is extremely effective.

4. The Haar-Fisz transform is usually a much better Gaussianiser and variance-stabiliser than the Anscombe transform, especially for lower intensities.

6.5 Poisson intensity estimation

Motivated by the excellent Gaussianisation, variance stabilisation and decorrelation properties of the Haar-Fisz transform demonstrated in the previous sections, we propose the following core algorithm for estimating the (possibly non-constant) intensity λ of a Poisson process:

[**A1**] Given the vector $\mathbf{v}$ of Poisson observations, preprocess it using the Haar-Fisz transformation to obtain $\mathcal{F}\mathbf{v}$.

[**A2**] Denoise $\mathcal{F}\mathbf{v}$ using *any* suitable ordinary wavelet denoising technique, appropriate for Gaussian noise (i.e. DWT — thresholding — inverse DWT). Denote the smoothed version of $\mathcal{F}\mathbf{v}$ by $\widehat{\mathcal{F}\lambda}$. We can optionally exploit the fact that the asymptotic variance of the noise is equal to one.

[**A3**] Perform the inverse Haar-Fisz transform to obtain $\mathcal{F}^{-1}(\widehat{\mathcal{F}\lambda})$ and take it to be the estimate of the intensity.

The following sections discuss several aspects of the above algorithm and compare its performance to a range of existing methods on a variety of test intensitites.

6.5.1 Methods for Poisson intensity estimation

Existing methods. As mentioned in Section 2.3.3, the Bayesian methods for Poisson intensity estimation due to Kolaczyk (1999a) and Timmermann & Nowak (1997, 1999) are currently state-of-the-art; see Besbeas *et al.* (2004). Our simulation study compared our technique with these Bayesian methods, as well as with the computationally intensive l_1-penalised likelihood technique of Sardy *et al.* (2004) and with a choice of methods based on the Anscombe transformation. To compare our technique with Kolaczyk (1999a) we used Eric Kolaczyk's *BMSMShrink* MATLAB software. As we did not have access to Timmermann and Nowak's software we exactly reproduced the simulation setup as in Timmermann & Nowak (1999) and compared our results to their Tables I and II. (Incidentally, the methods in Kolaczyk (1999a) and Timmermann & Nowak (1999) are very similar: the underlying Bayesian model is exactly the same, although the hyperparameter estimation is slightly different (Kolaczyk (2001), personal communication).)

Our method. The following describes the common features for our Poisson intensity estimation.

1. All our techniques always involve the Haar-Fisz transform, [**A1**], of the data, and the inverse Haar-Fisz transform, [**A3**].

2. In step [**A2**] of our algorithm the wavelet denoising technique may be of a translation invariant (TI) transform type; see Coifman & Donoho (1995). We refer to TI-denoising at this stage as "internal" cycle spinning (CS).

3. In step [**A2**] we could use any one of a number of wavelet families (e.g. multiwavelet, see Downie & Silverman (1998), complex-valued, see Lina (1997) etc.) for the denoising. In our simulations below we use Haar wavelets and Daubechies least-asymmetric wavelets of order 10; see Daubechies (1992).

4. Let $\mathcal{S}$ be the shift-by-one operator from Nason & Silverman (1995). The Haar-Fisz transform is not translation-equivariant since $\mathcal{F}\mathcal{S} \neq \mathcal{S}\mathcal{F}$. This

non-commutativity implies that it is beneficial to apply CS to the whole algorithm [**A1**]-[**A3**] even if [**A2**] uses a TI technique. We call this "external" CS.

Due to the particular type of nonlinearity of the Haar-Fisz transform there is no fast $O(N \log N)$ algorithm for the external CS. Therefore, we implement external CS by actually shifting the data before [**A1**], shifting back the estimate after [**A3**], and averaging over the estimates obtained through several different shifts.

For a data set of length N there are N possible shifts. However, through empirical investigation detailed in Section 6.5.3, we have found that 50 shifts are enough for data of length ≤ 1024. We postulate that using more shifts for longer data sets is likely to be beneficial.

Note that there is no point in doing external CS with the Anscombe transformation, $\mathcal{A}$, provided one has carried out internal CS, since Anscombe's transformation commutes with the shift operator: $\mathcal{A}\mathcal{S} = \mathcal{S}\mathcal{A}$.

The following list labels and describes the wavelet denoising methods that we choose to use in [**A2**]. In each case $\mathbf{F} \bowtie$ denotes the use of the Haar-Fisz transform and its inverse.

$\mathbf{F} \bowtie$ U: Universal hard thresholding from Donoho & Johnstone (1994) as implemented in *WaveThresh* (Nason (1998)) with default parameters (e.g. uses MAD variance estimation on all coefficients). 50 external cycle shifts.

$\mathbf{F} \bowtie$ CV: Cross-validation method from Nason (1996) as implemented in *WaveThresh* using default parameters but hard thresholding. 50 external cycle shifts.

$\mathbf{F} \bowtie$ BT: A variant of the greedy tree algorithm from Baraniuk (1999). 50 external cycle shifts.

Hybrids. We also looked at the performance of certain hybrid methods. These estimate the intensity by averaging the results of two of the above Haar-Fisz methods. Our main hybrid, **H:CV+BT**, combines $\mathbf{F} \bowtie \mathbf{CV}$ and $\mathbf{F} \bowtie \mathbf{BT}$. Note that hybrids can be easily formulated due to the large number of methods available for denoising Gaussian contaminated signals.

During our investigations we made use of several other denoisers including the eBayes procedure as described by Johnstone & Silverman (2003); universal hard threshold with internal cycle spinning; and hybrids of these with $\mathbf{F} \bowtie \mathbf{CV}$.

6.5.2 Simulation results for various test functions

The simulation setup in the first part of this section is the same as that described in Timmermann & Nowak (1999), who obtain two sets of intensity functions of length $N = 1024$ from the test functions from Donoho & Johnstone (1994). Each set is obtained by shifting and scaling to achieve (min,max) intensities of $(1/8, 8)$ and $(1/128, 128)$. The true intensity functions for the $(1/8, 8)$ case are shown as dashed lines in Figure 6.12.

The results for our methods are based on 100 independent simulations. The following list labels and provides details of other competing methods.

- BAY — Bayesian method developed in Timmermann & Nowak (1999), results quoted from the article, 25 independent simulations;

- BM — Bayesian *BMSMShrink* method developed in Kolaczyk (1999a), 100 independent simulations;

- L_1 P — l_1-penalised likelihood method from Sardy *et al.* (2004), results quoted from the article (after appropriate rescaling), 25 independent simulations. The authors only provide results for blocks and bumps;

- $\mathbf{A} \bowtie \mathbf{U}$ — method constructed in exactly the same way as the corresponding method $\mathbf{F} \bowtie \mathbf{U}$ except the (inverse) Anscombe transform was used instead of

Table 6.1: Normalised MISE values ($\times 10000$) for various existing techniques and our **F**$\bowtie$**U** and **H:CV+BT** methods using Haar wavelets and Daubechies' least asymmetric wavelets with 10 vanishing moments (LA10), on the test functions with peak intensities 8 and 128. The best results are indicated by a box.

Peak intensity = 8

			L_1 P	**A**$\bowtie$**U**		**F**$\bowtie$**U**		**H:CV+BT**
Intensity	BAY	BM	Haar	Haar	LA10	Haar	LA10	Haar
Doppler	154	146	*	218	121	201	$\boxed{99}$	159
Blocks	178	$\boxed{129}$	287	217	338	191	302	135
HeaviSine	52	46	*	98	63	68	$\boxed{40}$	64
Bumps	1475	1871	1557	3121	1579	2826	$\boxed{1268}$	2266

Peak intensity = 128

Doppler	26	20	*	28	$\boxed{12}$	29	$\boxed{12}$	23
Blocks	27	8	27	8	38	8	37	$\boxed{7}$
HeaviSine	$\boxed{7}$	$\boxed{7}$	*	9	$\boxed{7}$	9	$\boxed{7}$	9
Bumps	143	174	$\boxed{122}$	191	133	185	133	163

the (inverse) Haar-Fisz transform. 100 independent simulations.

The results reported in Table 6.1 are the MISE normalised by the squared l_2 norm of the true intensity vector, multiplied by 10000 and then rounded for clarity of presentation (this is exactly the same performance measure as in Timmermann & Nowak (1999) which is useful for comparability).

The results show that our **F**$\bowtie$**U** method with the LA10 wavelet outperforms the existing state-of-the-art methods especially for the lower intensity, except for the blocks function (as well as bumps for the higher intensity where our technique is outperformed by L_1 P by about 8%). The main reason why the performance of our method on blocks is less impressive is that a smooth wavelet is used in the Gaussian denoising step [**A2**]. As expected, the performance of **F**$\bowtie$**U** with the Haar wavelet is much better in this context, but still not as good as that of *BMSMShrink*, which is the best competitor for blocks. However, the hybrid method **H:CV+BT** with the Haar wavelet achieves performance comparable to *BMSMShrink*. We should emphasise here that our **F**$\bowtie$**U** method is far simpler to

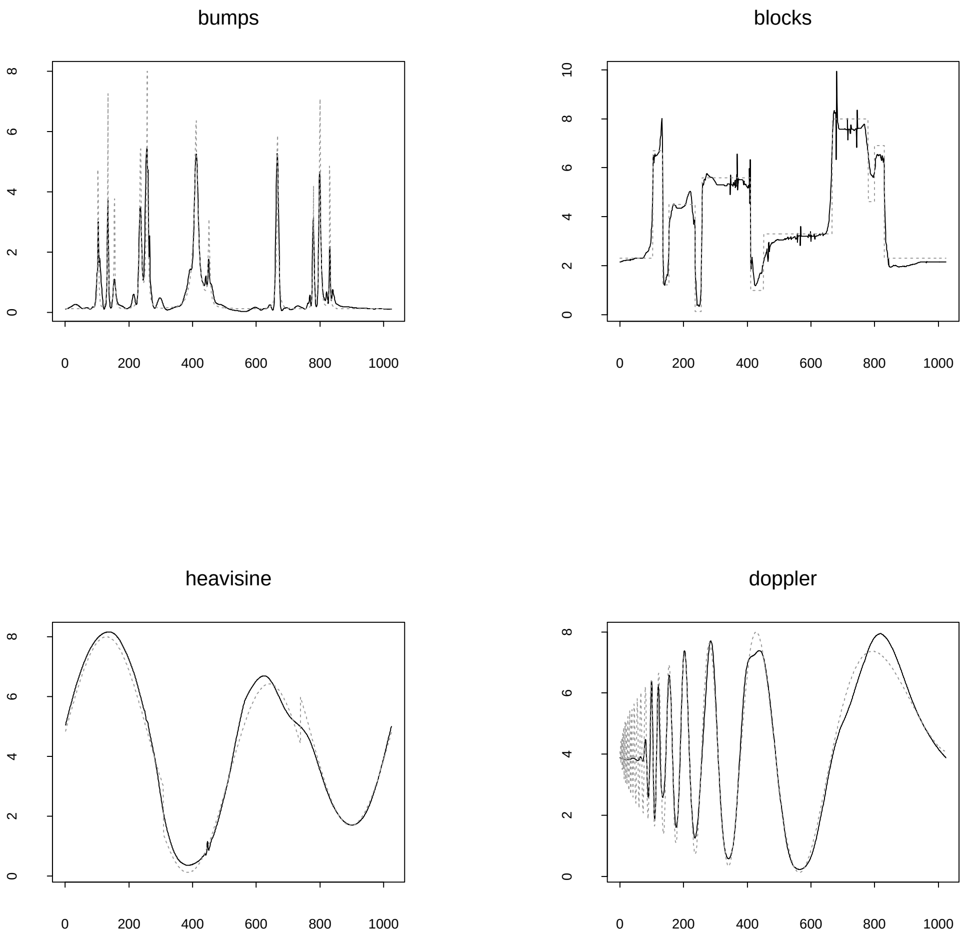

Figure 6.12: Selected estimates for the Donoho and Johnstone intensity functions (dashed, described in text). Each estimate gives an idea of "average" performance in that in each case its MISE is the closest to the median MISE obtained over 50 sample paths. The estimation method in each case was $\mathbf{F} \bowtie \mathbf{U}$ with Daubechies least-asymmetric wavelets with 10 vanishing moments except for blocks which used **H:CV+BT** with Haar wavelets.

implement than the current state-of-the-art techniques.

In Figure 6.12 the small spike in the heavisine function is not picked up well at intensity 8 but is almost always clearly estimated at intensity 128 (not shown). However, it should be said that the spike is almost completely obscured by noise in all realisations at intensity 8 so it would be extremely difficult for any method to detect it. We are impressed with the quality of the estimates using the new Haar-Fisz method, particularly with bumps and doppler. Also, the reconstruction of blocks, using the hybrid method **H:CV+BT**, is very accurate. Overall, it must be remembered that the reconstructions are usually going to be less impressive than the classical wavelet shrinkage problem where the test functions are contaminated with Gaussian noise with variance one.

To further investigate the performance of the methods on piecewise constant intensities we performed the following simulation study where the true intensity was the clipped blocks function of length $N = 1024$ shown on the left hand side of Figure 6.13. The clipped blocks intensity was obtained from the blocks function by setting all negative values to zero, scaling it so that the maximum intensity is 15.6 and then adding 3. We also examined the same intensity but scaled by factors of 1/6, 1/3 and 10/3. These scalings gave us a range of low and high intensity settings with large spreads of low intensity. The minimum and maximum intensities were, for each of these scalings: 3–18.6, 0.5–3.1, 1.0–6.2 and 10–62.

The simulation results reported in Table 6.2 are the MISE per bin: that is we computed the sum of the squared errors between our estimate and the true intensity, then divided by the number of bins (1024) and then took the mean over all 100 simulations.

Table 6.2 shows that at low to medium intensities *BMSMShrink* and **H:CV+BT** are competitive but at the higher 10/3 intensity our hybrid is about 10% better. The right hand figure in Figure 6.13 shows a particular sample reconstruction using the hybrid method **H:CV+BT**.

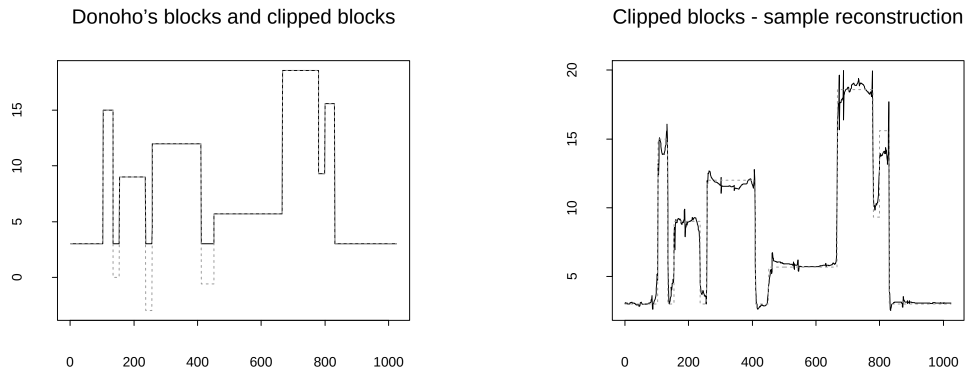

Figure 6.13: Left: Scaled and shifted blocks function, and its clipped version: clipped blocks. Right: The true intensity function (with scaling 1, dashed) and an estimate computed using our algorithm using hybrid method **H:CV+BT** whose MISE was closest to the median MISE obtained over 50 sample paths.

Table 6.2: MISE per bin ($\times 100$ and rounded) for clipped block intensity estimation using *BMSMShrink* and **H:CV+BT** as denoted in the text for a variety of intensity scalings.

	Scaling			
Method	1/6	1/3	1	10/3
BMSMShrink	9	20	61	191
H:CV+BT	9	20	61	171

6.5.3 Performance of Haar-Fisz methods as a function of the number of cycle shifts

In this section, we attempt to justify our choice of 50 external cycle shifts as the default number of CS recommended when estimating intensities of length ≤ 1024 (see Section 6.5.1). For the purpose of our simulation study, we selected the following intensities out of those considered in Section 6.5.2:

1. clipped blocks 1 — clipped blocks with scaling 1;
2. blocks 128 — blocks with maximum intensity 128;
3. bumps 128 — bumps with maximum intensity 8;
4. doppler 8 — doppler with maximum intensity 8;
5. heavisine 128 — heavisine with maximum intensity 128.

For each intensity, we examined the MISE depending on the number of cycle shifts (0–50), averaged over 10 simulated sample paths. This was done for both "short" (length 128), and "long" (length 1024) intensity vectors. Clipped blocks 1 and blocks 128 were estimated using $\mathbf{F}\bowtie\mathbf{CV}$ and $\mathbf{F}\bowtie\mathbf{BT}$ (components of the hybrid **H:CV+BT**) with Haar wavelets, whereas bumps 8, doppler 8 and heavisine 128 were estimated using $\mathbf{F}\bowtie\mathbf{U}$ with *DaubLeAsymm10* wavelets.

Successive rows in Figure 6.14 are results for clipped blocks 1 with $\mathbf{F}\bowtie\mathbf{CV}$, clipped blocks 1 with $\mathbf{F}\bowtie\mathbf{BT}$, blocks 128 with $\mathbf{F}\bowtie\mathbf{CV}$ and blocks 128 with $\mathbf{F}\bowtie\mathbf{BT}$. The left-hand column shows results for length 128, and the right-hand column — for length 1024.

Successive rows in Figure 6.15 show results for bumps 8, doppler 8 and heavisine 128. The meaning of columns is as above.

From the plots, it is evident that the MISE always stabilises after a small number of cycle-shifts. For clipped blocks 1, blocks 128 and bumps 8, the MISE stabilises after approximately 5 and 10 cycle-shifts (for short and long signals, respectively).

For doppler 8 and heavisine 128, the respective numbers of cycle-shifts are approximately 15 and 40.

Motivated by the above observation, we set the default value of the number of cycle-shifts to 50, bearing in mind that this number may need to be slightly increased in the case of extremely long (≥ 2048) signals.

6.6 Application to earthquake data

In this section, we analyse Northern Californian earthquake data, available from `http://quake.geo.berkeley.edu`. We analyse the time series N_k, $k = 1, \ldots, 1024$, where N_k is the number of earthquakes of magnitude 3.0 or more which occurred in the kth week, the last week under consideration being 29 November – 5 December 2000. The time series, imported into S-Plus, is plotted in Figure 6.16.

Our aim is to extract the intensity which underlies the realisation of this process. For the purposes of this example we shall use the *BMSMShrink* methodology of Kolaczyk (1999a) and our hybrid **H:CV+BT** method with Haar wavelets. The rationale for using **H:CV+BT** is that:

- it appears that the true earthquake intensity is highly non-regular and **H:CV+BT** with Haar wavelets worked the best on the blocks and clipped blocks simulation examples from Section 6.5.2;

- the earthquake data exhibits medium to high intensities and **H:CV+BT** was better than the other hybrids that we tried in this situation.

Figures 6.17 shows the intensity estimates obtained using *BMSMShrink* and **H:CV+BT** plotted on a log scale. (Due to the large peak at 274 weeks the original scale of 0-250 is not suitable for analysing the subtle differences between the estimates.) Visually the estimates are very similar however the **H:CV+BT** estimate is a little less variable. Although with this real data there is clearly

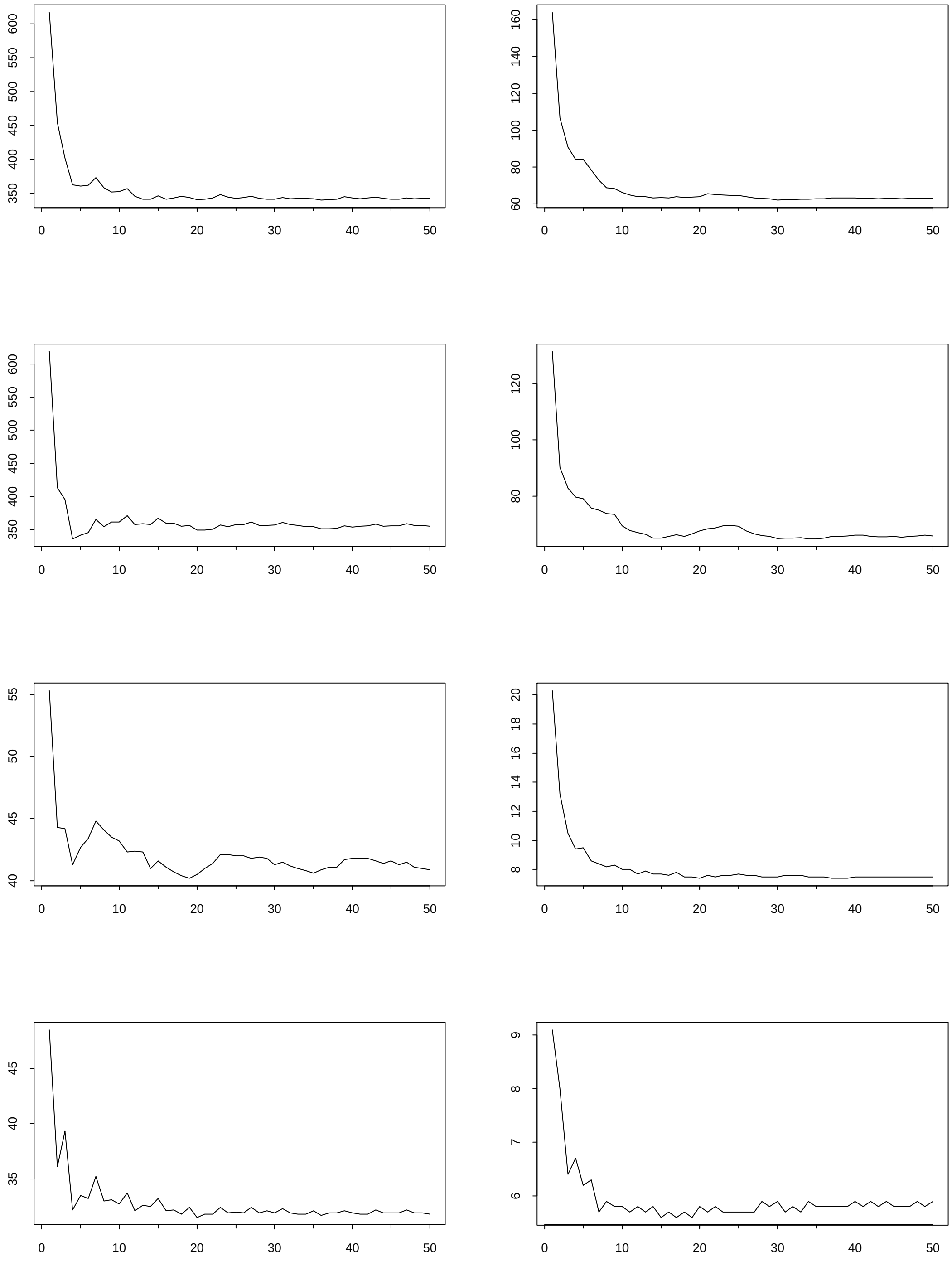

Figure 6.14: MISE against the number of shifts for clipped blocks 1 (top two rows) and blocks 128 (bottom two rows). See Section 6.5.3 for detailed description.

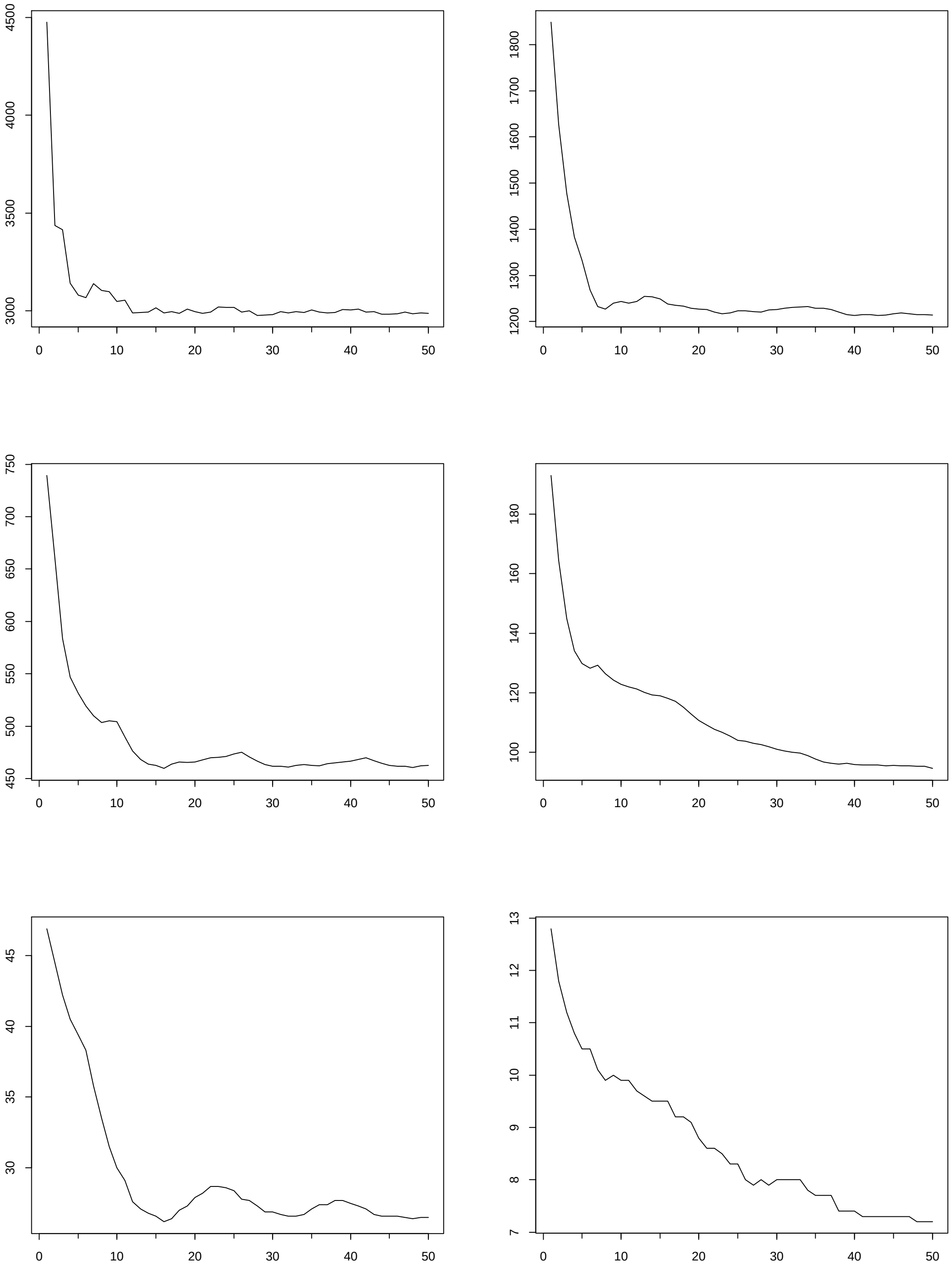

Figure 6.15: MISE against the number of shifts for bumps 8, doppler 8 and heavisine 128. See Section 6.5.3 for detailed description.

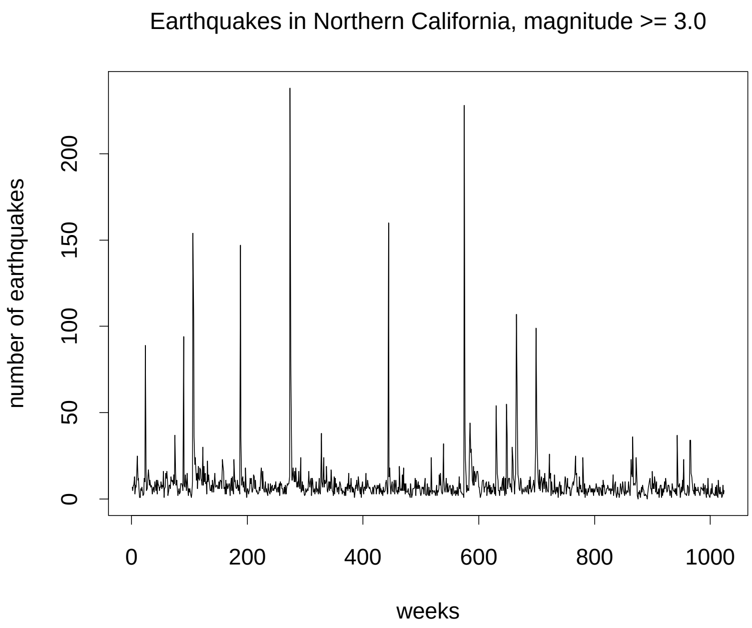

Figure 6.16: The number of earthquakes of magnitude ≥ 3.0 which occurred in Northern California in 1024 consecutive weeks, the last week being 29 Nov – 5 Dec 2000.

no right or wrong answer it is reassuring that they do give such similar visual results even though *BMSMShrink* and **H:CV+BT** are based on completely different philosophies.

6.7 Conclusion

In this chapter, we have described a new wavelet-based technique for bringing vectors of Poisson counts to normality with variance one. The technique, named the Haar-Fisz transformation, was applied to estimating the intensity of an inhomogeneous Poisson process, yielding a method whose performance was nearly always better than that of the current state-of-the-art.

For Poisson intensity estimation our methodology requires two components. The first, the Haar-Fisz transform, is very simple and easy to code. The second component can be any suitable Gaussian denoising procedure: we have used and compared a variety of wavelet methods ranging from the fast universal thresholding to

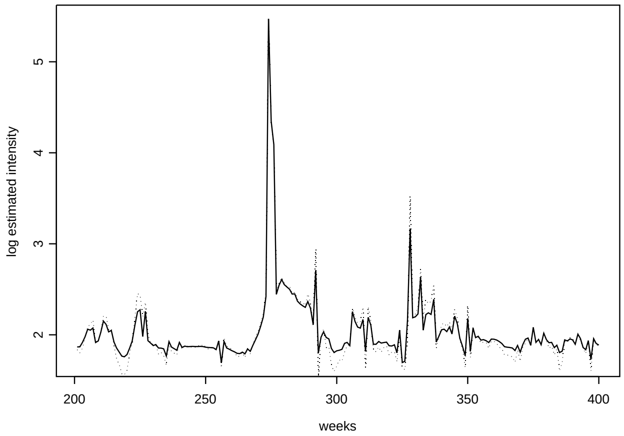

Figure 6.17: Intensity estimates for earthquake data for weeks 201 to 400. Dotted line is *BMSMShrink* estimate and solid is **H:CV+BT** estimate.

more complicated techniques such as cross-validation, "Baraniuk trees" and empirical Bayes. Since *any* Gaussian denoiser can be used, the Haar-Fisz algorithm can only improve as the field develops.

If computational speed is not an issue, and little is known about the smoothness of the true intensity, we recommend that several denoisers be used and a hybrid averaging all of their results, with optional full cycle-spinning, be considered. However, if speed is important then there is an issue over which one denoiser should be chosen: not all denoisers are appropriate for all types of intensity as our earlier simulations confirmed. Our recommendation is that if one suspects the intensity is piecewise constant then one should use Haar wavelets and a hybrid method such as **H:CV+BT**; otherwise, we strongly recommend the use of $\mathbf{F} \bowtie \mathbf{U}$ with a smooth wavelet.

We believe that one of the reasons why the performance of the Haar-Fisz algorithm is so good is due to the non-commutativity of the Haar-Fisz and shift operators, which enables meaningful cycle spinning. Also, the Fisz transform itself is a more effective normaliser than Anscombe.

The S-Plus routines implementing the algorithm, as well as the data set, are included on the associated CD.

Chapter 7

Conclusions and future directions

In this thesis, we have considered some wavelet-based and wavelet-related approaches to selected problems arising in two important branches of statistics: time series analysis and Poisson regression. In this chapter, we briefly summarise the main contributions made in Chapters 3 – 6 and then move on to discuss possible directions for future research.

In Chapter 3, we considered several theoretical and computational aspects of forecasting Gaussian Locally Stationary Wavelet (LSW) processes by means of the linear predictor. As direct MSPE minimisation would have required knowledge of unidentifiable parameters, we proposed to compute the prediction coefficients by *approximate* MSPE minimisation. We identified conditions under which the approximation was valid but found that one of the them was overly restrictive. We circumvented that theoretical difficulty by introducing a slight modification to the definition of an LSW process (the new class was labelled "LSW_2"). The minimisation of the MSPE led to the generalisation of the Yule-Walker equations: the stability of the system was analysed. We also found that "sparse" LSW_2 processes were ill conditioned for forecasting. To conclude the theoretical part of the work, we derived a generalisation of Kolmogorov's formula for one-step MSPE in the LSW_2 framework. In practice, the entries of the prediction matrix have to be estimated, and we studied the behaviour of the first two moments of our (multiscale) estimators. We then proposed a complete forecasting algorithm where the choice of the arising nuisance parameters was performed by adaptive forecasting. The algorithm was successfully applied to the forecasting of a meteorological time series. Our work in Chapter 3 provided an answer to the interesting question of whether and how wavelets could be useful in forecasting non-stationary time series.

In the less technical Chapter 4, we attempted to model financial log-returns in the Gaussian LSW framework. We first extended the LSW model to include the time-modulated white noise (TMWN) as a special case (we labelled the new class "LSW_3"). Then, we used theory to show that the LSW_3 model was able to pick up the most commonly observed stylised facts of financial time series. We proposed a generic automatic algorithm for estimating the time-varying wavelet spectrum of log-returns with guaranteed nonnegativity, and used simulation to demonstrate the excellent performance of its two particular implementations. An exploratory data analysis of the FTSE 100 series showed that its second order structure was changing over time, and that Haar wavelets were ideally suited for the modelling of that series. Finally, we showed by simulation that financial log-returns could be successfully forecast in the LSW_3 model.

In Chapter 5, we proposed a fast multiscale method, called the Haar-Fisz transform, for stabilising the variance of the wavelet periodogram (WP) in the Gaussian LSW model and bringing its distribution closer to normality. To be able to analyse the Haar-Fisz transform from a theoretical point of view, we stated and proved a Functional Central Limit Theorem (FCLT) for the WP. We then formulated and proved the Gaussianising, variance stabilising and decorrelating properties of the transform in a certain asymptotic regime. We showed by simulation that the Haar-Fisz transform is a far better Gaussianiser than the classical variance stabilising log transform. We then proposed a denoising method for the WP, which consisted of three basic steps: take the Haar-Fisz transform, denoise the transformed WP using a method suitable for signals contaminated with stationary Gaussian noise, and then take the inverse Haar-Fisz transform. We assessed the performance of the method using simulation and showed that it outperformed an existing competitor in most of the cases. We concluded the chapter by using the Haar-Fisz methodology to perform a local variance analysis of the Dow Jones index: the analysis showed that the series could be modelled as Gaussian.

In Chapter 6, we introduced a Haar-Fisz transform for sequences of Poisson counts, again with the aim of stabilising their variance as well as bringing their distribution close to Gaussianity. We proved that if the underlying Poisson intensity was constant, then the Haar-Fisz transformed vector was asymptotically normal with variance one, and its elements were uncorrelated. We used simulation to show that the Gaussianising, variance stabilising and decorrelating properties of the Haar-Fisz transform also held for time-varying intensities. Also, we demonstrated that the Haar-Fisz transform was a more effective Gaussianiser and variance stabiliser than the usual Anscombe square-root transform. Then, we proposed a method for estimating Poisson intensities, based on the Haar-Fisz transform. Our technique outperformed state-of-the-art competitors in most of the cases; occasionally, its performance was slightly inferior but comparable. We applied our estimation technique to the well-known Northern California earthquake data: visually, our method gave similar results to the current state of the art.

We conclude this thesis by considering a few possible avenues for further research. The adaptive forecasting algorithm of Section 3.5 merits further investigation: in particular, it would be interesting to examine the dependence of the nuisance parameters on their initial values, to further robustify the algorithm, and to investigate its theoretical properties (e.g. its asymptotic behaviour as $T \to \infty$).

Also, we suspect that by using the MSPE criterion in Chapter 3, we do not fully exploit the potential of the LSW model in forecasting non-stationary time series. Indeed, observe that the prediction matrix in Section 3.1 is similar to the variance-covariance matrix of the process, which means that the wavelet spectrum (the main quantity of interest in LSW modelling) is only used indirectly there, through the local autocovariance. There arises a question of whether a tractable (multiscale) prediction criterion can be formulated which would make more direct use of the wavelet spectrum.

As was shown in Section 3.2, sparse LSW processes are “ill conditioned” for forecasting. It would be interesting to investigate whether a multiscale time series model could be formulated whereby processes which are represented sparsely were, in an appropriate sense, “good” for forecasting. As for financial time series modelling (Chapter 4), the problem of volatility forecasting in the LSW model could make an interesting research project, as could the issue of using “skewed” wavelets and/or non-Gaussian innovations.

Furthermore, it remains to be investigated whether and how the Haar-Fisz transform for the wavelet periodogram (a) can be extended to processes with a discontinuous spectral structure; (b) can be used for testing for time series stationarity. Also, it would be desirable to theoretically quantify the variance stabilising property of the Haar-Fisz transform for $M = \log_2(T)$ (see Section 5.5.2).

The problem of the choice of the primary resolution when the noise is correlated is still an open question, badly neglected in the wavelet literature. Indeed, we are unaware of any automatic method for performing this selection. It would clearly make an exciting and potentially very rewarding research project.

Also, it would be of interest to establish a theoretical proof of the Gaussianising, variance stabilising and decorrelating properties of the Haar-Fisz transform for non-stationary Poisson signals. Finally, an exciting possibility for future research would be to investigate how the Haar-Fisz transform can be used for Gaussianising other distributions, not only χ^2 or Poisson. Recall that in the χ^2 case, we use the Fisz transform with exponent 1, and in the Poisson case — with exponent 1/2. In the case of other distributions, we could attempt to estimate the suitable exponent, or indeed a suitable function of the Haar smooth coefficient, from the data.

Bibliography

Abramovich, F., & Benjamini, Y. 1996. Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, **22**, 351–361.

Abramovich, F., & Silverman, B. W. 1998. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, **85**, 115–129.

Abramovich, F., Bailey, T. C., & Sapatinas, T. 2000. Wavelet analysis and its statistical applications. *J. Roy. Statist. Soc. Ser. D*, **49**, 1–29.

Altman, N. S. 1990. Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, **85**, 749–759.

Anscombe, F. J. 1948. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246–254.

Antoniadis, A., & Sapatinas, T. 2001. Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, **88**, 805–820.

Antoniadis, A., Grégoire, G., & Nason, G. P. 1999. Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Stat. Soc. Ser. B*, **61**, 63–84.

Audit, B., Bacry, E., Muzy, J. F., & Arneodo, A. 2002. Wavelet-based estimators of scaling behaviour. *IEEE Transactions on Information Theory*, **48**, 2938–2954.

Aussem, A., & Murtagh, F. 2001. Web traffic demand forecasting using wavelet-based multiscale decomposition. *International Journal of Intelligent Systems*, **16**, 215–236.

BIBLIOGRAPHY

Averkamp, R., & Houdré, C. 2003. Wavelet thresholding for non-necessarily Gaussian noise: idealism. *Ann. Stat.*, **31**, 110–151.

Baraniuk, R. G. 1999. Optimal tree approximation with wavelets. *Pages 206–214 of:* Unser, M. A., Aldroubi, A., & Laine, A. F. (eds), *Wavelet Applications in Signal and Image Processing VII.* Proceedings of SPIE, vol. 3813. SPIE.

Barber, S., & Nason, G. P. 2003. Real nonparametric regression using complex wavelets. *Technical Report 03:06, Department of Mathematics, University of Bristol, UK.*

Battaglia, F. 1979. Some extensions in the evolutionary spectral analysis of a stochastic process. *Boll. Un. Mat. Ital. B (5)*, **16**, 1154–1166.

Bera, A. K., & Higgins, M. L. 1993. ARCH models: properties, estimation and testing. *J. Economic Surveys*, **7**, 305–366.

Besbeas, P., De Feis, I., & Sapatinas, T. 2004. A comparative simulation study of wavelet shrinkage estimators for Poisson counts. *International Statistical Review*, **72**.

Bilen, C., & Huzurbazar, S. 2002. Wavelet-based detection of outliers in time series. *Journal of Computational and Graphical Statistics*, **11**, 311–327.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, **31**, 307–327.

Bougerol, P., & Picard, N. 1992. Stationarity of GARCH processes and of some nonnegative time series. *J. Econometrics*, **52**, 115–127.

Brillinger, D. R. 1998. Some wavelet analyses of point process data. *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers, 1997*, **2**, 1087–1091.

Brockwell, P. J., & Davis, R. A. 1987. *Time Series: Theory and Methods.* Springer.

Brooks, C. 1997. Linear and non-linear (non-)forecastability of high-frequency exchange rates. *J. Forecasting*, **16**, 125–145.

Cai, T., & Silverman, B. W. 2001. Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā Ser. B*, **63**, 127–148. Special issue on wavelets.

Calvet, L., & Fisher, A. 2001. Forecasting multifractal volatility. *J. Econometrics*, **105**, 27–58.

Candès, E. J., & Donoho, D. L. 2001. Curvelets and curvilinear integrals. *J. Approx. Theory*, **113**, 59–90.

Chatfield, C. 1996. *The Analysis of Time Series: An Introduction.* 5 edn. Chapman & Hall.

Chatfield, C. 2000. *Time Series Forecasting.* Chapman & Hall.

Chiann, C., & Morettin, P. 1999. A wavelet analysis for time series. *J. Nonparametr. Statist.*, **10**, 1–46.

Chui, C. 1992. *An Introduction To Wavelets.* Academic Press.

Cohen, A. 2003. *Numerical Analysis of Wavelet Methods.* Studies in Mathematics and Its Applications, vol. 32. Elsevier.

Cohen, A., Daubechies, I., & Feauveau, J. 1992. Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, **45**, 485–560.

Cohen, A., Daubechies, I., & Vial, P. 1993. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, **1**, 54–81.

Coifman, R. R., & Donoho, D. L. 1995. Translation-invariant de-noising. *Technical Report, Statistics Department, Stanford University.*

BIBLIOGRAPHY

Coifman, R. R., & Wickerhauser, M. V. 1992. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, **38**, 713–718.

Coifman, R. R., Meyer, Y., Quake, S., & Wickerhauser, M. V. 1989. Signal processing and compression with wave packets. *In:* Meyer, Y. (ed), *Proceedings of the International Conference on Wavelets, Marseilles.* Paris: Masson.

Cooley, J. W., & Tukey, O. W. 1965. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, **19**, 297–301.

Cox, D. R., Hinkley, D. V., & Barndorff-Nielsen, O. E. (eds). 1996. *Time Series Models in Econometrics, Finance and Other Fields.* Monographs on Statistics and Applied Probability, vol. 65. Chapman & Hall.

Cristan, A. C., & Walden, A. T. 2002. Multitaper power spectrum estimation and thresholding: Wavelet packets versus wavelets. *IEEE Trans. Sig. Proc.*, **50**, 2976–2986.

Dahlhaus, R. 1996a. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. *In:* Robinson, P. M., & Rosenblatt, M. (eds), *Athens Conference on Applied Probability and Time Series Analysis, vol. 2.* New York: Springer-Verlag.

Dahlhaus, R. 1996b. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, **62**, 139–168.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Ann. Stat.*, **25**, 1–37.

Dahlhaus, R., & Neumann, M. H. 2001. Locally adaptive fitting of semiparametric models to nonstationary time series. *Stoch. Proc. Appl.*, **91**, 277–308.

Dahlhaus, R., Neumann, M. H., & von Sachs, R. 1999. Non-linear wavelet estimation of time-varying autoregressive processes. *Bernoulli*, **5**, 873–906.

Daubechies, I. 1992. *Ten Lectures on Wavelets.* Philadelphia, Pa.: SIAM.

Davidson, J. 1994. *Stochastic Limit Theory.* Oxford University Press.

Davidson, J. 2003. Moment and memory properties of linear conditional heteroscedasticity models. *Journal of Business and Economic Statistics.*

Donoho, D., Johnstone, I., Kerkyacharian, G., & Picard, D. 1996. Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

Donoho, D. L. 1993. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. *Pages 173–205 of: Proc. Sympos. Appl. Math.* Amer. Math. Soc.

Donoho, D. L. 1995. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, **2**, 101–126.

Donoho, D. L. 2000. Orthonormal ridgelets and linear singularities. *SIAM J. Math. Anal.*, **31**, 1062–1099.

Donoho, D. L., & Huo, X. 2002. Beamlets and multiscale image analysis. *Pages 149–196 of: Multiscale and multiresolution methods.* Lect. Notes Comput. Sci. Eng., vol. 20. Berlin: Springer.

Donoho, D. L., & Johnstone, I. M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. L., & Johnstone, I. M. 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Downie, T., & Silverman, B. W. 1998. The discrete multiple wavelet transform and thresholding methods. *IEEE Transactions on Signal Processing*, **46**, 2558–2561.

Engle, R. F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

BIBLIOGRAPHY

Farge, M., Kevlahan, N., Perrier, V., & Schneider, K. 1999. Turbulence analysis, modelling and computing using wavelets. *Pages 117–200 of: Wavelets in physics.* Cambridge: Cambridge Univ. Press.

Fisz, M. 1955. The limiting distribution of a function of two independent random variables and its statistical application. *Colloquium Mathematicum*, **3**, 138–146.

Fryzlewicz, P., & Nason, G. P. 2004. A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics,* to appear.

Fryzlewicz, P., Van Bellegem, S., & von Sachs, R. 2003. Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Stat. Math.*, **55**, 737–764.

Gao, H-Y. 1997. Choice of thresholds for wavelet shrinkage estimate of the spectrum. *J. Time Ser. Anal.*, **18**, 231–251.

Gençay, R., Selçuk, F., & Whitcher, B. 2001. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press.

Gencay, R., Selcuk, F., & Whitcher, B. 2001. Scaling properties of foreign exchange volatility. *Physica A*, **289**, 249–266.

Geronimo, J. S., Hardin, D. P., & Massopust, P. R. 1994. Fractal functions and wavelet expansions based on several scaling functions. *Journal of Approximation Theory*, **78**, 373–401.

Geva, A. B. 1998. ScaleNet — Multiscale neural-network architecture for time series prediction. *IEEE Transactions on Neural Networks*, **9**, 1471–1482.

Green, P. J., & Silverman, B. W. 1994. *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall.

Grillenzoni, C. 2000. Time-varying parameter prediction. *Ann. Inst. Statist. Math.*, **52**, 108–122.

Haar, A. 1910. Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.*, **69**, 331–371.

Hall, P., & Patil, P. 1995. Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.*, **23**, 905–928.

Härdle, W., Spokoiny, V. G., & Teyssière, G. 2000. Adaptive estimation for a time inhomogeneous stochastic volatility model. *SFB Discussion Paper No. 6/00. Berlin: Humboldt University.*

Hastie, T. J., & Tibshirani, R. J. 1990. *Generalized Additive Models.* Chapman & Hall.

Hee, Y. Y., Chong, F. L., & Zhong, B. L. 2002. A hierarchical evolutionary algorithm for constructing and training wavelet networks. *Neural Computing & Applications*, **10**, 357–366.

Herrick, D., Nason, G. P., & Silverman, B. W. 2001. Some new methods for wavelet density estimation. *Sankhyā Ser. A*, **63**, 394–411. Special issue on wavelets.

Hoffmann, M. 1999. On nonparametric estimation in nonlinear AR(1) models. *Statistics & Probability Letters*, **44**, 29–45.

Hong, Y. M., & Lee, J. 2001. One-sided testing for ARCH effects using wavelets. *Econometric Theory*, **17**, 1051–1081.

Ikeda, Y., & Tokinaga, S. 1999. Evaluation of stock option prices by using the prediction of fractal time-series. *Journal of the Operations Research Society of Japan*, **42**, 18–31.

Jaffard, S., Meyer, Y., & Ryan, R. D. 2001. *Wavelets: Tools for Science & Technology.* Philadelphia, Pa.: SIAM.

Johnstone, I. M., & Silverman, B. W. 1997. Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B*, **59**, 319–351.

Johnstone, I. M., & Silverman, B. W. 2003. Empirical Bayes selection of wavelet thresholds. *Technical Report 02:17, Department of Mathematics, University of Bristol, UK.*

Kerkyacharian, G., & Picard, D. 1996. Estimating nonquadratic functionals of a density using Haar wavelets. *Ann. Statist.*, **24**, 485–507.

Kim, W. 1998. Econometric analysis of locally stationary time series models. *Manuscript, Yale University.*

Kokoszka, P., & Leipus, R. 2000. Change-point estimation in ARCH models. *Bernoulli*, **6**, 513–539.

Kolaczyk, E. D. 1997. Non-parametric estimation of Gamma-ray burst intensities using Haar wavelets. *The Astrophysical Journal*, **483**, 340–349.

Kolaczyk, E. D. 1999a. Bayesian multiscale models for Poisson processes. *Journal of the American Statistical Association*, **94**, 920–933.

Kolaczyk, E. D. 1999b. Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statistica Sinica*, **9**, 119–135.

Kress, R. 1991. *Numerical Analysis.* Springer.

Lawton, W. 1993. Applications of complex valued wavelet transforms to subband decomposition. *IEEE Trans. Sig. Proc*, **41**, 3566–3568.

Ledolter, J. 1981. Recursive estimation and adaptive forecasting in ARIMA models with time varying coefficients. *Pages 449–471 of: Applied time series analysis, II (Tulsa, Okla., 1980).* New York: Academic Press.

Lee, J., & Hong, Y. M. 2001. Testing for serial correlation of unknown form using wavelet methods. *Econometric Theory*, **17**, 386–423.

Leung, M. T., Daouk, H., & Chen, A.-S. 2000. Forecasting stock indices: a comparison of classification and level estimation models. *Int. J. Forecasting*, **16**, 173–190.

Li, T. H., & Hinich, M. J. 2002. A filter bank approach for modeling and forecasting seasonal patterns. *Technometrics*, **44**, 1–14.

Li, Y. A., & Xie, Z. J. 1997. The wavelet detection of hidden periodicities in time series. *Statistics & Probability Letters*, **35**, 9–23.

Lina, J-M. 1997. Image processing with complex Daubechies wavelets. *Journal of Mathematical Imaging and Vision*, **7**, 211–223.

Lina, J-M., & Mayrand, M. 1995. Complex Daubechies wavelets. *Applied and Computational Harmonic Analysis*, **2**, 219–229.

Mackenzie, D. 2001. Wavelets: seeing the forest and the trees. *Available from* `http://www.beyonddiscovery.org/content/view.article.asp?a=1952`.

Maddala, G. S., & Rao, C. R. (eds). 1996. *Statistical Methods in Finance*. Handbook of Statistics, vol. 14. Elsevier.

Mallat, S. 1989a. Multiresolution approximations and wavelet orthonormal bases of $\mathbb{L}_2(\mathbb{R})$. *Trans. Amer. Math. Soc*, **315**, 69–87.

Mallat, S. 1989b. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Mallat, S. 1998. *A Wavelet Tour of Signal Processing*. Academic Press.

Mallat, S., Papanicolaou, G., & Zhang, Z. 1998. Adaptive covariance estimation of locally stationary processes. *Ann. Stat.*, **26**, 1–47.

Masuda, N., & Okabe, Y. 2001. Time series analysis with wavelet coefficients. *Japan Journal of Industrial and Applied Mathematics*, **18**, 131–160.

Mélard, G., & Herteleer-De Schutter, A. 1989. Contributions to evolutionary spectral theory. *J. Time Ser. Anal.*, **10**, 41–63.

Meyer, Y. 1992. *Wavelets and Operators.* Cambridge University Press.

Mikosch, T., & Starica, C. 2003. Change of structure in financial data, long range dependence and GARCH modelling. *The Review of Economics and Statistics,* to appear.

Milidiu, R. L., Machado, R. J., & Renteria, R. P. 1999. Time-series forecasting through wavelets transformation and a mixture of expert models. *Neurocomputing*, **28**, 145–156.

Morlet, J., Arens, G., Fourgeau, E., & Giard, D. 1982. Wave propagation and sampling theory. *Geophysics*, **47**, 203–236.

Murty, K. G. 1988. *Linear Complementarity, Linear and Nonlinear Programming.* Internet Edition, available from `http://ioe.engin.umich.edu/people/fac/books/murty/linear_complementarity_webbook/`.

Nason, G. P. 1996. Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, **58**, 463–479.

Nason, G. P. 1998. *WaveThresh3 Software.* Available from `http://www.stats.bris.ac.uk/~wavethresh/`.

Nason, G. P., & Sapatinas, T. 2002. Wavelet packet transfer function modelling of nonstationary time series. *Statistics and Computing*, **12**, 45–56.

Nason, G. P., & Silverman, B. W. 1994. The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.

Nason, G. P., & Silverman, B. W. 1995. The stationary wavelet transform and some statistical applications. *Pages 281–300 of:* Antoniadis, A., & Oppenheim, G. (eds), *Lecture Notes in Statistics, vol. 103.* Springer-Verlag.

Nason, G. P., & von Sachs, R. 1999. Wavelets in time series analysis. *Philosophical Transactions of the Royal Society of London (Series A)*, **357**, 2511–2526.

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

Neumann, M., & von Sachs, R. 2000. A wavelet-based test for stationarity. *J. Time Ser. Anal.*, **21**, 597–613.

Neumann, M. H. 1996. Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *Journal of Time Series Analysis*, **17**, 601–633.

Neumann, M. H., & von Sachs, R. 1997. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, **25**, 38–77.

Nowak, R. D., & Baraniuk, R. G. 1999. Wavelet domain filtering for photon imaging systems. *IEEE Transactions on Image Processing*, **8**, 666–678.

Ogden, R. T., & Parzen, E. 1996. Change-point approach to data analytic wavelet thresholding. *Statist. Comput.*, **6**, 93–99.

Ombao, H., Raz, J., von Sachs, R., & Malow, B. 2001a. Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Stat. Assoc.*, **96**, 543–560.

Ombao, H., Raz, J., von Sachs, R., & Guo, W. 2002. The SLEX model of a non-stationary random process. *Ann. Inst. Statist. Math*, **54**, 171–200.

Ombao, H. C., Raz, J. A., Strawderman, R. L., & von Sachs, R. 2001b. A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika*, **88**, 1186–1192.

Opsomer, J.D., Wang, Y., & Yang, Y. 2001. Nonparametric regression with correlated errors. *Statistical Science*, **16**, 134–153.

Patil, P. N., & Wood, A. T. 2004. Counting process intensity estimation by orthogonal wavelet methods. *Bernoulli,* to appear.

Penev, S., & Dechevsky, L. 1997. On non-negative wavelet-based density estimators. *J. Nonparametr. Statist.*, **7**, 365–394.

Pensky, M. 1999. Estimation of a smooth density function using Meyer-type wavelets. *Statist. Decisions*, **17**, 111–123.

Pensky, M. 2002. Density deconvolution based on wavelets with bounded supports. *Statist. Probab. Lett.*, **56**, 261–269.

Pensky, M., & Vidakovic, B. 1999. Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, **27**, 2033–2053.

Percival, D. B., & Walden, A. T. 2000. *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

Pesquet, J. C., Krim, H., & Carfantan, H. 1996. Time-invariant orthonormal wavelet representations. *IEEE Trans. Sig. Proc.*, **44**, 1964–1970.

Philander, S. 1990. *El Niño, La Niña and the southern oscillation.* San Diego: Academic Press.

Pinheiro, A., & Vidakovic, B. 1997. Estimating the square root of a density via compactly supported wavelets. *Comput. Statist. Data Anal.*, **25**, 399–415.

Prakasa Rao, B. L. S. 1999. Estimation of the integrated squared density derivatives by wavelets. *Bull. Inform. Cybernet.*, **31**, 47–65.

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

Priestley, M. B. 1981. *Spectral Analysis and Time Series.* Academic Press.

Ramsey, J. 1999. The contribution of wavelets to the analysis of economic and financial data. *Phil. Trans. Roy. Soc. London A*, **357**, 2593–2606.

Ramsey, J. 2002. Wavelets in economics and finance: past and future. *Studies in Nonlinear Dynamics and Econometrics*, **6**.

Rao, T. S., & Indukumar, K. C. 1996. Spectral and wavelet methods for the analysis of nonlinear and nonstationary time series. *J. Franklin Inst. — Eng. Appl. Math.*, **333B**, 425–452.

Ruskai, M. B. (ed). 1992. *Wavelets and Their Applications.* Jones and Barlett books in mathematics. Jones and Berlett.

Sakiyama, K. 2002. Some statistical applications for locally stationary processes. *Sci. Math. Jpn.*, **56**, 231–250.

Sardy, S., Antoniadis, A., & Tseng, P. 2004. Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics,* to appear.

Schneider, K., & Farge, M. 2001. Computing and analyzing turbulent flows using wavelets. *Pages 181–216 of: Wavelet transforms and time-frequency signal analysis.* Appl. Numer. Harmon. Anal. Boston, MA: Birkhäuser Boston.

Serroukh, A., Walden, A. T., & Percival, C. B. 2000. Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *Journal of the American Statistical Association*, **95**, 184–196.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics.* Springer.

Soltani, S. 2002. On the use of wavelet decomposition for time series prediction. *Neurocomputing*, **48**, 267–277.

Soltani, S., Boichu, D., Simard, P., & Canu, S. 2000. The long-term memory prediction by multiscale decomposition. *Signal Processing*, **80**, 2195–2205.

Struzik, Z. R. 2001. Wavelet methods in (financial) time series processing. *Physica A*, **296**, 307–319.

Sweldens, W. 1996. The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, **3**, 186–200.

Swift, R. 2000. The evolutionary spectra of a harmonizable process. *J. Appl. Statist. Sci.*, **9**, 265–275.

Taylor, S. J. 1986. *Modelling Financial Time Series*. Chichester: Wiley.

Timmermann, K. E., & Nowak, R. D. 1997. Multiscale Bayesian estimation of Poisson intensities. *Pages 95–90 of: Proceedings of the Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA: IEEE Computer Press.

Timmermann, K. E., & Nowak, R. D. 1999. Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory*, **45**, 846–862.

Triebel, H. 1983. *Theory of Function Spaces*. Basel: Birkhäuser Verlag.

Truong, Y. K., & Patil, P. N. 2001. Asymptotics for wavelet based estimates of piecewise smooth regression for stationary time series. *Annals of the Institute of Statistical Mathematics*, **53**, 159–178.

Vanreas, E., Jansen, M., & Bultheel, A. 2002. Stabilized wavelet transforms for non-equispaced data smoothing. *Signal Processing*, **82**, 1979–1990.

Vidakovic, B. 1999. *Statistical Modeling by Wavelets*. New York: Wiley.

von Sachs, R., & MacGibbon, B. 2000. Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scand. J. Statist.*, **27**, 475–499.

von Sachs, R., & Schneider, K. 1996. Wavelet smoothing of evolutionary spectra by nonlinear thresholding. *Applied and Computational Harmonic Analysis*, **3**, 268–282.

Walden, A. T., & Serroukh, A. 2002. Wavelet analysis of matrix-valued time series. *Proc. Roy. Soc. London Ser. A*, **458**, 157–179.

Walden, A. T., Percival, D. B., & McCoy, E. J. 1998. Spectrum estimation by wavelet thresholding of multitaper estimators. *IEEE Trans. Sig. Proc.*, **48**, 3153–3165.

Walter, G., & Shen, X. 1999. Deconvolution using Meyer wavelets. *J. Integral Equations Appl.*, **11**, 515–534.

Wand, M. P., & Jones, M. C. 1994. *Kernel Smoothing.* Chapman & Hall.

Wang, Y. Z., Cavanaugh, J. E., & Song, C. Y. 2001. Self-similarity index estimation via wavelets for locally self-similar processes. *J. Statist. Plan. Infer.*, **99**, 91–110.

Wasserman, P. 1993. *Advanced Methods in Neural Computing.* New York: Van Nostrand Reinhold.

West, M., & Harrison, J. 1997. *Bayesian Forecasting and Dynamic Models.* 2 edn. Springer.

Whitcher, B. 2001. Simulating Gaussian processes with unbounded spectra. *Journal of Computational and Graphical Statistics*, **10**, 112–134.

Whitcher, B., Byers, S. D., Guttorp, P., & Percival, D. B. 2002. Testing for homogeneity of variance in time series: Long memory, wavelets and the Nile River. *Water Resources Research*, **38**.

Wong, H., Ip, W., & Li, Y. 2001. Detection of jumps by wavelets in a heteroscedastic autoregressive model. *Stat. Probab. Lett.*, **52**, 365–372.

Wong, H., Ip, W. C., Xie, Z. J., & Lui, X. L. 2003. Modelling and forecasting by wavelets, and the application to exchange rates. *Journal of Applied Statistics*, **30**, 537–553.

Zhang, B. L., & Dong, Z. Y. 2001. An adaptive neural-wavelet model for short term load forecasting. *Electric Power Systems Research*, **59**, 121–129.

Zhang, B. L., Coggins, R., Jabri, M. A., Dersch, D., & Flower, B. 2001a. Multiresolution forecasting for futures trading using wavelet decompositions. *IEEE Transactions on Neural Networks*, **12**, 765–775.

Zhang, G. P., Patuwo, B. E., & Hu, M. Y. 2001b. A simulation study of artificial neural networks for nonlinear time series forecasting. *Computer and Operations Research*, **28**, 381–396.

Zheng, T. X., Girgis, A. A., & Makram, E. B. 2000. A hybrid wavelet-Kalman filter method for load forecasting. *Electric Power Systems Research*, **54**, 11–17.

Zheng, Y. J., Lin, Z. P., & Tay, D. B. H. 2001. State-dependent vector hybrid linear and nonlinear ARMA modelling: Applications. *Circuits Systems and Signal Processing*, **20**, 575–597.