Modelling and forecasting financial log-returns as locally stationary wavelet processes

Piotr Fryźlewicz *

February 17, 2004

Abstract

In this article, we model financial log-return series in the Locally Stationary Wavelet (LSW) framework proposed by Nason et al. (2000). We slightly alter the LSW setup to include time-modulated white noise (TMWN) as a special case. We show that the LSW model, being linear and non-stationary, adequately captures the most commonly observed stylised facts.

Furthermore, we propose a new method for estimating time-varying second order quantities in the LSW model, and provide an exploratory analysis of the daily FTSE 100 series using the LSW toolbox. The example shows that the dependence structure of FTSE 100 varies over time, and that the LSW model is particularly well suited for modelling this series.

Finally, by considering daily returns on the DJIA index, we demonstrate that financial log-returns can be successfully forecast in the LSW framework.

Keywords: non-decimated wavelets, wavelet periodogram, stylised facts, non-stationarity, adaptive forecasting.

1 Introduction

Financial log-return series, be it stock index returns or exchange rates, often exhibit the following well-known properties:

1. The sample mean of the series is close to zero.

*Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ, UK. Tel.: +44 (0)20 75895111 ext. 58600, fax: +44 (0)20 75948517, email: p.fryzlewicz@imperial.ac.uk.

2. The marginal distribution is roughly symmetric (or only slightly skewed), has a peak at zero, and is heavy-tailed.

3. (a) The sample autocorrelations of the series are "small" at almost all lags.

   (b) The sample autocorrelations of the absolute values and squares of the series are significant for a large number of lags.

4. Volatility is "clustered", i.e. days of either large or small movements are followed by days of similar characteristics.

Clearly, to capture the above stylised facts, one needs to look beyond the stationary linear framework, and to preserve stationarity, a huge number of non-linear models have been proposed. Among them, two branches seem to be the most popular: the family of Autoregressive Conditionally Heteroscedastic (ARCH) models, and Stochastic Volatility (SV) models. The ARCH model was proposed by Engle (1982) and Generalised ARCH (GARCH), its most popular extension — independently by Bollerslev (1986) and Taylor (1986). The SV model was suggested by Taylor (1986) as an alternative to ARCH-type modelling. Literature on both these families is massive, some recent recommendable monographs are Cox et al. (1996), Maddala and Rao (1996) and Fan and Yao (2003). Giraitis et al. (2003) is a recent review article on various aspects of ARCH modelling.

Even though the assumption of stationarity is attractive from the estimation point of view, some authors point out that the stylised facts listed above can be better accounted for by the possible non-stationarity of the series, see for example Mikosch and Starica (2003), Kokoszka and Leipus (2000) (who look at the detection of change points in the ARCH model) or Härdle et al. (2000) (who introduce a time-varying SV model and look at the adaptive estimation of its parameters). An attractive asymptotic framework for modelling non-stationary time series was proposed by Dahlhaus (1996), who developed a theory of Locally Stationary Fourier (LSF) processes. Some attempts have been made to apply Dahlhaus's theory in finance: e.g. Kim (1998) provides various statistical analyses of financial and macroeconomic data in the LSF framework (however, he does not consider forecasting).

In this paper, we also adopt the "locally stationary" approach, and model log-returns in the Locally Stationary Wavelet (LSW) framework of Nason et al. (2000). Essentially, this implies that the log-return series is composed of discrete wavelet vectors at various scales, rather than localised Fourier functions at various frequencies like in the LSF theory. There are two main

initial motivations behind using wavelets, rather than Fourier functions, as building blocks in the model. Firstly, many authors observe that various economic factors operate at different time scales (see for example Calvet and Fisher (2001)), and wavelets are a commonly used tool in the analysis of multiscale phenomena (see Vidakovic (1999) for an overview of wavelet applications in statistics). Secondly, Fryźlewicz et al. (2003) developed a working algorithm for forecasting LSW processes, which we are able to take advantage of and use in our context. To our knowledge, no such algorithm has been proposed and tested for the LSF model.

On the other hand, our approach differs from GARCH/SV modelling in that both GARCH and SV are nonlinear, but often stationary models, whereas the LSW model is linear, but only locally stationary.

The aims of this paper are:

- to argue that the LSW model can accurately account for the most commonly observed stylised facts,
- to propose a new (suitable for log-returns) method of estimating time-varying second order quantities in the LSW model,
- to demonstrate the attractiveness of the LSW framework as a tool for the exploratory analysis of log-return data,
- to demonstrate that log-returns can be successfully forecast using the adaptive forecasting algorithm of Fryźlewicz et al. (2003).

We do not aim to demonstrate that the LSW methodology is uniformly superior to any other method of analysis of financial data. Instead, we propose to treat it as yet another tool in the toolbox, particularly useful for forecasting or exploring the local structure of log-return series.

All the theoretical results in the paper have been obtained for Gaussian LSW processes. The following sections show that most stylised facts can be explained using this simple class; however, we emphasize here that due to the simplicity of the (linear) LSW model, analogous theoretical results can easily be obtained for other noise distributions.

Even though the examples provided in the paper use stock index returns only, the LSW methodology can also in principle be applied to other instruments, such as shares or exchange rates.

The paper is organised as follows: in Section 2, 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 3, we recall the LSW model and show that Gaussian TMWN is a special case of an LSW process. In Section 4, we provide theoretical evidence that LSW processes can capture most of the stylised facts listed at the beginning of this section. In Section 5, we introduce a new (suitable for log-returns) estimation approach for LSW processes, and demonstrate its superiority to the general method of Nason et al. (2000). In Section 6, we provide an interesting example of exploratory data analysis using the LSW model. Finally, in Section 7, we apply the adaptive forecasting algorithm of Fryźlewicz et al. (2003) to log-returns, and provide a comparison with forecasts based on GARCH modelling. Section 8 concludes the paper.

2 Motivation

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 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 3, we show that Gaussian TMWN is a special case of an LSW process.

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 paper, is

`http://bossa.pl/notowania/daneatech/metastock`

(page in Polish). X_t is plotted in the top left subfigure of Figure 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 5 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 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 paper, 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 and 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 recall the LSW model and show that Gaussian TMWN is a special case of a general LSW process. In Section 6, we come back to the example of FTSE 100 and model this series in the general LSW framework. We show that, in this way, more local features of the FTSE 100 data can be picked up.

3 The model

Definition 3.1 (Nason et al. (2000)) *An LSW process* $\{X_{t,T}\}_{t=0,1,\ldots,T-1}$, $T = 2^J \geq 2$, *is defined as*

$$X_{t,T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} \omega_{j,k;T} \psi_{j,k-t} \xi_{j,k}, \tag{1}$$

where

1. *The parameters j and k denote scale and location, respectively.*
2. *The random innovations $\xi_{j,k}$ have mean 0 and $\mathbb{E}(\xi_{j,k}, \xi_{j',k'}) = \delta_{j,j'}\delta_{k,k'}$, where $\delta_{m,n} = 1$ if $m = n$ and 0 otherwise.*
3. *The amplitudes $\omega_{j,k;T}$ are real constants and $\forall\, j \leq -1 \quad \exists\, W_j : [0,1) \to \mathbb{R}$ such that W_j is Lipschitz with parameter L_j and*

$$\sum_{j=-\infty}^{-1} W_j^2 < \infty \tag{2}$$

$$\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty \tag{3}$$

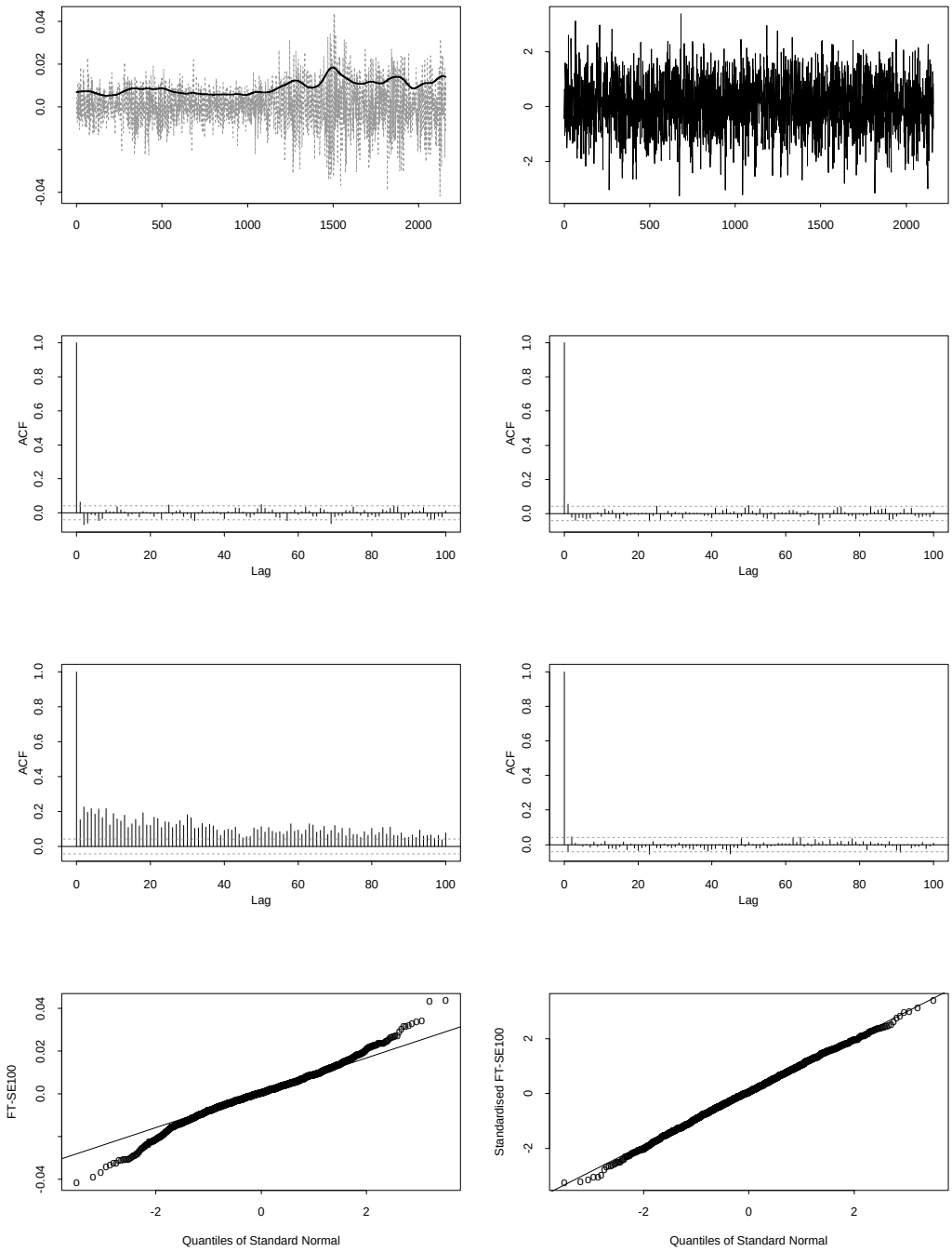

Figure 1: Left-hand column, from top to bottom: X_t with $\hat{\sigma}_t$ superimposed, acf of X_t, acf of X_t^2, normal QQ plot of X_t. Right-hand column, from top to bottom: Z_t, acf of Z_t, acf of Z_t^2, normal QQ plot of Z_t. See Section 2 for a discussion.

$$\exists \{C_j\}_{j \leq -1} \forall T \sup_{k=0,1,\ldots,T-1} |\omega_{j,k;T} - W_j(k/T)| \leq C_j/T \tag{4}$$

$$\sum_j C_j < \infty. \tag{5}$$

4. *The vectors* $\psi_j = (\psi_{j,0}, \psi_{j,1}, \ldots, \psi_{j,\mathcal{L}_j})$, $j = -1, -2, \ldots, -J$, $\mathcal{L}_j \sim M2^{-j}$, $M \geq 1$, *are discrete wavelets.*

For a mathematical introduction to wavelets, the reader is referred to Daubechies (1992), and for an overview of their applications in statistics, to Vidakovic (1999). By way of example, we recall the simplest discrete wavelet system: the Haar wavelets. They are defined by

$$\psi_{j,k} = 2^{j/2}\mathbb{I}_{\{0,1,\ldots,2^{-j-1}-1\}}(k) - 2^{j/2}\mathbb{I}_{\{2^{-j-1},\ldots,2^{-j}-1\}}(k) \quad \text{for} \quad j = -1, -2, \ldots \quad \text{and} \quad k \in \mathbb{Z}. \tag{6}$$

In the above, $j = -1$ is the finest scale, $j = -2$ is the second finest scale, etc. We can think of (1) as an analogue of the traditional Cramér representation for stationary processes, where a process is a linear combination of Fourier basis functions. Here, sines and cosines have been replaced by more "localised" wavelets, therefore potentially allowing a successful modelling of non-stationary series.

Condition (4) means that, for each j, the sequence $\{\omega_{j,k;T}\}_k$ is "closer and closer" (as $T \to \infty$) to a Lipschitz function $W_j(z)$ defined on the interval $[0, 1)$. The "slow" evolution of the sequence $\{\omega_{j,k;T}\}_k$ makes it possible to establish an asymptotic framework which enables effective estimation in the model. The rescaled time setup implies that letting $T \to \infty$ does *not* mean obtaining information about the future; instead, it means obtaining more and more information about the local structure of the process. See also Nason et al. (2000) and Dahlhaus (1996) for further discussion of the concept of rescaled time.

Nason et al. (2000) define the *evolutionary wavelet spectrum* of $X_{t,T}$ as $S_j(z) = W_j(z)^2$. For stationary processes, the spectrum is independent of time: we have $\omega^2_{j,k;T} = W_j^2 = S_j$.

In the classical theory the autocovariance function and the spectrum of a stationary process are Fourier transforms of each other, and an analogous link can be established between the evolutionary wavelet spectrum and the *local autocovariance*. The finite-sample local autocovariance in the LSW model is defined as

$$c_T(z, \tau) = \text{cov}(X_{[zT],T}, X_{[zT]+\tau,T}).$$

Nason et al. (2000) show that $c_T(z,\tau)$ has an asymptotic limit as $T \to \infty$. Indeed, define the *autocorrelation wavelets* as

$$\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{j,k}\psi_{j,k-\tau},$$

and define the local autocovariance as

$$c(z,\tau) = \sum_{j=-\infty}^{-1} S_j(z)\Psi_j(\tau). \tag{7}$$

Proposition 3.1 (Nason et al. (2000)) *With the asymptotics of Definition 3.1,* $\|c_T - c\|_{L_\infty} = O(T^{-1})$.

The above Proposition says that the local autocovariance is the "autocorrelation wavelet" transform of the evolutionary wavelet spectrum.

Theorem 1 in Nason et al. (2000) states that the evolutionary wavelet spectrum (and, therefore, the local autocovariance), are uniquely defined given an LSW process. There is a one-to-one correspondence between $\{S_j(z)\}_j$ and $\{c(z,\tau)\}_\tau$, and an inverse formula to (7) can be derived.

The local variance is denoted by $\sigma^2(z) := c(z,0)$.

Before looking at two important examples of LSW processes, we quote the following useful lemma from Fryźlewicz et al. (2003).

Lemma 3.1 (Fryźlewicz et al. (2003)) *Let* $\{\Psi_j\}_j$ *be the autocorrelation wavelets constructed from Daubechies' compactly supported wavelets of an arbitrary degree of smoothness (this includes Haar wavelets as a special case). We have*

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \delta_0(\tau),$$

where $\delta_0(k) = 1$ *if* $k = 0$ *and* 0 *otherwise.*

Example 1 (white noise). By Lemma 3.1, if $X_{t,T} = Z_t$ where Z_t is i.i.d. $N(0,1)$, then $X_{t,T}$ is LSW with $S_j = 2^j$.

Example 2 (time-modulated white noise). Suppose that $X_{t,T} = \sigma(t/T)Z_t$ with Z_t i.i.d. $N(0,1)$. By Lemma 3.1, if $X_{t,T}$ was LSW, we would have to have $S_j(z) = \sigma^2(z)2^j$. However, we would then have $L_j = L2^{j/2}$, where L is the Lipschitz constant for $\sigma(z)$, and that would violate condition (3). This shows that, without modifications, the LSW model cannot accommodate

time-modulated white noise, which, as we saw in Section 2, is an essential basic model for financial log-returns. To remedy this unwelcome situation, we slightly alter the definition of an LSW process.

Definition 3.2 *An LSW process* $\{X_{t,T}\}_{t=0,1,\ldots,T-1}$, $T = 2^J \geq 2$, *is defined as*

$$X_{t,T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} \omega_{j,k;T} \psi_{j,k-t} \xi_{j,k}, \tag{8}$$

where

1. *The parameters* j *and* k *denote scale and location, respectively.*

2. *The random innovations* $\xi_{j,k}$ *have mean 0 and* $\mathbb{E}(\xi_{j,k}, \xi_{j',k'}) = \delta_{j,j'}\delta_{k,k'}$.

3. *The amplitudes* $\omega_{j,k;T}$ *are real constants and* $\forall\, j \leq -1 \quad \exists\, W_j \in C([0,1))$ *such that* $S_j := W_j^2$ *is Lipschitz with parameter* L_j *and*

$$\sup_{z,j} S_j(z) 2^{-j} = D < \infty. \tag{9}$$

$$\sum_{j=-J}^{-1} 2^{-j} L_j = O(\log(T)) \tag{10}$$

$$\exists\, \{C_j\}_{j \leq -1} \forall\, T \sup_{k=0,1,\ldots,T-1} |\omega_{j,k;T} - W_j(k/T)| \leq C_j/T \tag{11}$$

$$\sum_j C_j < \infty \tag{12}$$

4. *The vectors* $\psi_j = (\psi_{j,0}, \psi_{j,1}, \ldots, \psi_{j,\mathcal{L}_j})$, $j = -1, -2, \ldots, -J$, $\mathcal{L}_j \sim M 2^{-j}$, $M \geq 1$, *are discrete wavelets.*

It is now easy to verify that Gaussian TMWN with σ Lipschitz satisfies the assumptions of Definition 3.2 with $\omega_{j,k;T} = W_j(k/T) = \sigma(k/T) 2^{j/2}$.

Under the assumptions of Definition 3.2, 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 3.2 *With the asymptotics of Definition 3.2,* $\|c_T - c\|_{L_\infty} = O(T^{-1} \log(T))$.

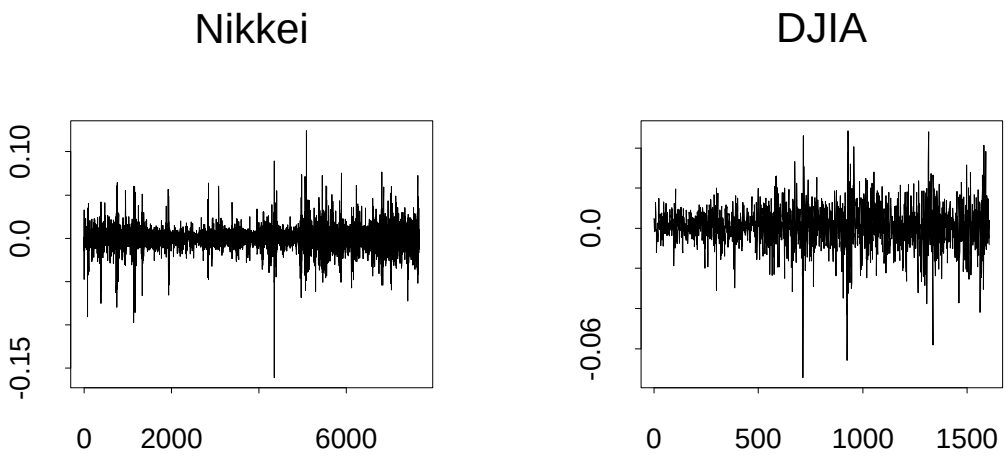

Figure 2: Left-hand plot: log-returns on daily closing values of Nikkei (5/6 Jan 1970 – 11-14 May 2001). Right-hand plot: log-returns on daily closing values of DJIA (3-4 Jan 1995 – 10/11 May 2001).

Throughout the rest of the paper, we will work with Definition 3.2, rather than Definition 3.1.

Innovations $\xi_{j,k}$. So far, we have considered $\xi_{j,k}$ i.i.d. $N(0,1)$. Fryzlewicz and Nason (2003) argue that Gaussian innovations in the LSW model account surprisingly well even for extreme events such as those present in the Nikkei index (left-hand plot in Figure 2) or the DJIA index (right-hand plot in Figure 2). Nevertheless, we believe that, occasionally, other distributions of $\xi_{j,k}$ may need to be used: for example, 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 this article 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 paper, we assume $\mathbb{E}(X_{t,T}) = 0$ (as is obvious from Definition 3.2). 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 and MacGibbon (2000).

4 Explanation of the stylised facts

In this section, we demonstrate that Gaussian LSW 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.

Heavy tails of the "marginal" distribution. In this paragraph, we consider the sample second moment and the sample kurtosis:

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

For stationary Gaussian series, 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.

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

For the purpose of this paragraph, denote the first summand in the above formula by A^2. Obviously, $A^2 = 0$ iff the variance of $X_{t,T}$ is constant. Therefore, for a non-constant $\sigma^2(z)$, we will have

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

The above formula 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.

Negligible sample autocorrelations. As in Mikosch and Starica (2003), we consider the sample autocovariance function

$$\gamma_T(\mathbf{X}, h) = \frac{1}{T} \sum_{t=0}^{T-1-h} X_{t,T} X_{t+h,T} - \left(\frac{1}{T} \sum_{t=0}^{T-1} X_{t,T} \right)^2, \tag{13}$$

and the sample autocorrelation function

$$\rho_T(\mathbf{X}, h) = \frac{\gamma_T(\mathbf{X}, h)}{\gamma_T(\mathbf{X}, 0)}. \tag{14}$$

Also, we define the *scalogram*:

$$\bar{S}_i = \frac{1}{T} \sum_{t=0}^{T-1} S_i \left(\frac{t}{T} \right). \tag{15}$$

The following proposition holds:

Proposition 4.1 *For an LSW process $X_{t,T}$, we have*

$$\mathbb{E}(\gamma_T(\mathbf{X}, h)) = \sum_i \bar{S}_i \Psi_i(h). \tag{16}$$

By Lemma 3.1, the above quantity will be "close" to $C\delta_0(h)$ if $\bar{S}_i$ is "close" to $C2^i$. The examples provided in Section 6 demonstrate that it is indeed often the case. This would explain the often negligible sample autocorrelations of log-returns.

Non-negligible sample autocorrelations of the squares. The following proposition holds:

Proposition 4.2 *For an LSW process $X_{t,T}$, we have*

$$\mathbf{E}(\gamma_T(\mathbf{X^2}, h)) = \frac{1}{T} \sum_{t=0}^{T-1} \left(\sigma^2 \left(\frac{t}{T} \right) - \frac{1}{T} \sum_{s=0}^{T-1} \sigma^2 \left(\frac{s}{T} \right) \right)^2 + \frac{2}{T} \sum_{t=0}^{T-1} c^2 \left(\frac{t}{T}, h \right) + O\left(\frac{h + \log(T)}{T} \right). \tag{17}$$

For the purpose of this paragraph, denote the first summand of formula (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 and 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.

Clustering of volatility. The "clustering of volatility" or, in other words, a "slowly varying local variance" is indeed one of the features of LSW modelling.

5 Estimation

To estimate the spectrum, Nason et al. (2000) use the *wavelet periodogram*:

$$I_{j,p} = \left| \sum_t X_{t,T} \psi_{j,p-t} \right|^2, \quad j = -1, -2, \ldots, -J, \quad p = 0, 1, \ldots, T-1, \tag{18}$$

where ψ is the same wavelet family which is used to build $X_{t,T}$. In our altered setup of Definition 3.2, we will also use the statistic defined by (18). The following proposition holds.

Proposition 5.1 *Let $X_{t,T}$ satisfy Definition 3.2. We have*

$$\mathbb{E} I_{j,p} = \sum_{i=-\infty}^{-1} S_i \left(\frac{p}{T}\right) A_{ij} + O\left(\frac{2^{-j}\log(T)}{T}\right), \tag{19}$$

where A is defined as in Nason et al. (2000):

$$A_{ij} = \sum_\tau \Psi_i(\tau)\Psi_j(\tau).$$

In addition, if $X_{t,T}$ is Gaussian, then

$$\mathrm{Var}(I_{j,p}) = 2\left(\sum_i S_i\left(\frac{p}{T}\right)A_{ij}\right)^2 + O\left(\frac{2^{-j}\log(T)}{T}\right). \tag{20}$$

The form of the remainder in (19) suggests that the estimator is more accurate for finer scales, i.e. $-j \ll J$.

Formula (19) suggests the following method of estimating the spectrum: we solve the system of equations

$$I_{j,k} = \sum_i \hat{S}_{i,k}A_{ij} \tag{21}$$

to obtain an approximately unbiased estimator $\hat{S}_{j,k}$ of the spectrum $S_j(k/T)$ (see Nason et al. (2000) for details).

However, formula (20) shows that the wavelet periodogram is not a consistent estimator and needs to be smoothed to obtain consistency. We can either first solve (21), and then smooth $\hat{S}_{j,k}$, or first smooth $I_{j,k}$, and then solve (21). Following Nason et al. (2000), we prefer the latter option, as it is often easier to work out the distributional properties of $I_{j,k}$ than those of $\hat{S}_{j,k}$, and therefore it is easier to justify the choice of smoothing parameters for $I_{j,k}$.

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 $\mathbb{E}I_{j,k}/(\mathrm{Var}I_{j,k})^{1/2} \approx 1/\sqrt{2}$), and also to a significant amount of autocorrelation present in $I_{j,k}$. Nason et al. (2000) propose an adaptive wavelet denoising method, which, however, does not perform particularly well when applied to financial log-returns: this will be demonstrated in Section 5.4.

In Section 5.1, we propose an alternative general methodology for smoothing the wavelet periodogram. Section 5.2 looks at two specific methods of smoothing, and Section 5.3 deals with inverting (21) in an approximate manner to ensure the nonnegativity of the estimated spectrum.

5.1 General algorithm

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

orthogonality of the *decimated* wavelets, the sequence

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

as well as the sequence

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

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

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

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

$$I_{j,i}, I_{j,i+2^{-j}}, \ldots, I_{j,i+T-2^{-j}}, \quad i = 0, 1, \ldots, 2^{-j} - 1 \tag{22}$$

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 better numerically (on simulated data) and are visually more appealing (on real data) than that proposed by Nason et al. (2000). 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 $I_{j,k}$ as independent.

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

1. Fix j.

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

$$I_{j,i}, I_{j,i+2^{-j}}, \ldots, I_{j,i+T-2^{-j}}$$

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. (2001) can be applied, as we are also dealing with a sample of independent gamma variates, like in periodogram smoothing. In Section 5.2, we explore two other methods in which the smoothing parameter is chosen by CV.

3. Interpolate the smoothed sequence at all the points $0, 1, \ldots, T-1$ (e.g. using linear interpolation). Denote the interpolated smoothed sequence by

$$\tilde{I}^{(i)}_{j,0}, \tilde{I}^{(i)}_{j,1}, \ldots, \tilde{I}^{(i)}_{j,T-1}.$$

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

$$\hat{I}_{j,k} = \sum_{i=0}^{2^{-j}-1} \tilde{I}^{(i)}_{j,k}$$

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

The estimates $\hat{I}_{j,k}$ can now be substituted into the systems of linear equations

$$\hat{I}_{j,k} = \sum_i \hat{S}_{i,k} A_{ij}. \tag{23}$$

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.

5.2 Smoothing the decimated periodogram

In step 2 of the algorithm of Section 5.1, we apply a smoothing procedure to the decimated subsequences of the wavelet periodogram. In this section, we advertise the use of two smoothing methods: *cubic B-splines* (see Hastie and Tibshirani (1990) for details) and *translation-invariant linear wavelet smoothing* (see Donoho and Coifman (1995) and Nason and Silverman (1995)).

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

- The method performs well (see Section 5.4).
- 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 (19) and (20)), 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 5.4.

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

- The method performs well (see Section 5.4).
- The only smoothing parameter to be chosen is the "primary resolution", above which all the wavelet coefficients are set to zero, see Nason and Silverman (1995). As there are only $\log_2(T)$ primary resolution levels to choose from, the choice is potentially easier than the choice of bandwidth in kernel smoothing. We perform the selection by "leave-half-out" cross-validation as in Nason (1996), except that we choose the primary resolution rather than the threshold.
- The method is fast, as in practice we choose the primary resolution for the wavelet periodogram at the finest scale $j = -1$, and then use the same primary resolution for all the coarser scales $j = -2, -3, \ldots, -J$.

Adaptive methods allowing the detection of abrupt changes in the wavelet periodogram would

clearly be an attractive alternative. The method proposed by Fryzlewicz and Nason (2003), based on the Haar-Fisz transform, seems to be particularly promising in this context.

5.3 Estimating the spectrum with guaranteed nonnegativity

The evolutionary wavelet spectrum $S_j(z)$ is a nonnegative quantity so it would also be desirable if $\hat{S}_{j,k}$ was guaranteed to be nonnegative. This can be achieved, for example, by replacing the system of equations (23) by a Linear Complementarity Problem (LCP; see e.g. Murty (1988)):

$$\begin{aligned} A\hat{\mathbf{S}}_k &\geq \hat{\mathbf{I}}_k \\ \hat{\mathbf{S}}_k &\geq 0 \\ \left(A\hat{\mathbf{S}}_k - \hat{\mathbf{I}}_k\right)\hat{\mathbf{S}}_k &= 0. \end{aligned}$$

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

Let $\hat{S}_{j,k}^{\mathrm{LCP}}$ denote the estimate of $S_j(k/T)$ obtained using the LCP formulation, and $\hat{S}_{j,k}^{\mathrm{INV}}$ — using the simple inversion of formula (23). By (7), we estimate the local variance $\sigma^2(k/t)$ in each case by

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

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 $\sigma^2(k/T)$:

$$\hat{S}_{j,k} = \hat{\sigma}^2(k/T)^{(\mathrm{INV})} \frac{\hat{S}_{j,k}^{\mathrm{LCP}}}{\hat{\sigma}^2(k/T)^{(\mathrm{LCP})}} \tag{24}$$

$$\hat{\sigma}^2(k/T) = \sum_{j=-J}^{-1} \hat{S}_{j,k}. \tag{25}$$

As explained in Sections 5.1 and 5.2, $\hat{S}_{j,k}$ depends on the method used for smoothing the wavelet periodogram. The next section compares the estimators based on cubic B-splines and linear wavelet denoising to that proposed by Nason et al. (2000).

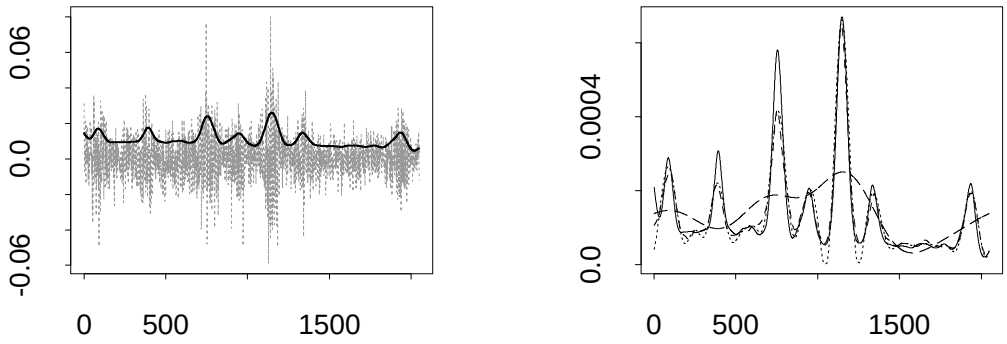

Figure 3: Left-hand plot: sample path from Gaussian TMWN model with time-varying standard deviation superimposed. Right-hand plot: time-varying variance (solid), its estimate using splines (dot-dashed), its estimate using linear wavelet scheme (dotted), and its estimate using the method of Nason et al. (2000) with default parameters (dashed).

5.4 Estimation — numerical results

The left-hand plot in Figure 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 (25)). The right-hand plot shows the time-varying variance (solid line), the estimate obtained using spline smoothing (dot-dashed line), the estimate obtained using translation-invariant linear wavelet smoothing with Daubechies' least asymmetric wavelet with 10 vanishing moments (dotted line), and the estimate obtained using the adaptive method of Nason et al. (2000) with default parameters (dashed line).

While the estimates using our two methods almost coincide with each other and with the true time-varying variance, 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 our two methods and the method by Nason et al. (2000) basing on 25 simulated sample paths. We used two criterion

	default	splines	TI wavelets
mean of d_{σ^2}	1197	189	162
mean of d_S	460	227	232

Table 1: Values of the criterion functions averaged over 25 simulations. "Default" is the method of Nason et al. (2000), "splines" is our method using spline smoothing and "TI wavelets" is our method using translation-invariant linear wavelet scheme.

functions — one for the Haar spectrum:

$$d_S(\hat{S}, S) = \left[\frac{10^{11}}{T} \sum_{i=-J}^{-1} \sum_{t=0}^{T-1} \left(\hat{S}_i \left(\frac{t}{T} \right) - S_i \left(\frac{t}{T} \right) \right)^2 \right], \tag{26}$$

and the other for the variance:

$$d_{\sigma^2}(\hat{\sigma}^2, \sigma^2) = \left[\frac{10^{11}}{T} \sum_{t=0}^{T-1} \left(\hat{\sigma}^2 \left(\frac{t}{T} \right) - \sigma^2 \left(\frac{t}{T} \right) \right)^2 \right]. \tag{27}$$

The values in Table 1 confirm our observation that the two estimators in which the choice of the smoothing parameter is performed by cross-validation give very similar results.

6 Exploratory data analysis

In this section, we look at two examples of data analysis using the LSW methodology (the examples are related to each other). The first one uses the Haar scalogram (see formula (15)), and the other — the full evolutionary Haar wavelet spectrum.

Scalogram. In this example, we compute the Haar scalogram for four series:

- $X_{t,T}$: the last 1024 observations of the artificial simulated Gaussian TMWN of Figure 3,
- $F_{t,T}$: the last 1024 observations of the FTSE 100 series of Figure 1,
- $N_{t,T}$: the last 1024 observations of the Nikkei series of Figure 2,
- $D_{t,T}$: the last 1024 observations of the Dow Jones IA series of Figure 2.

Figure 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

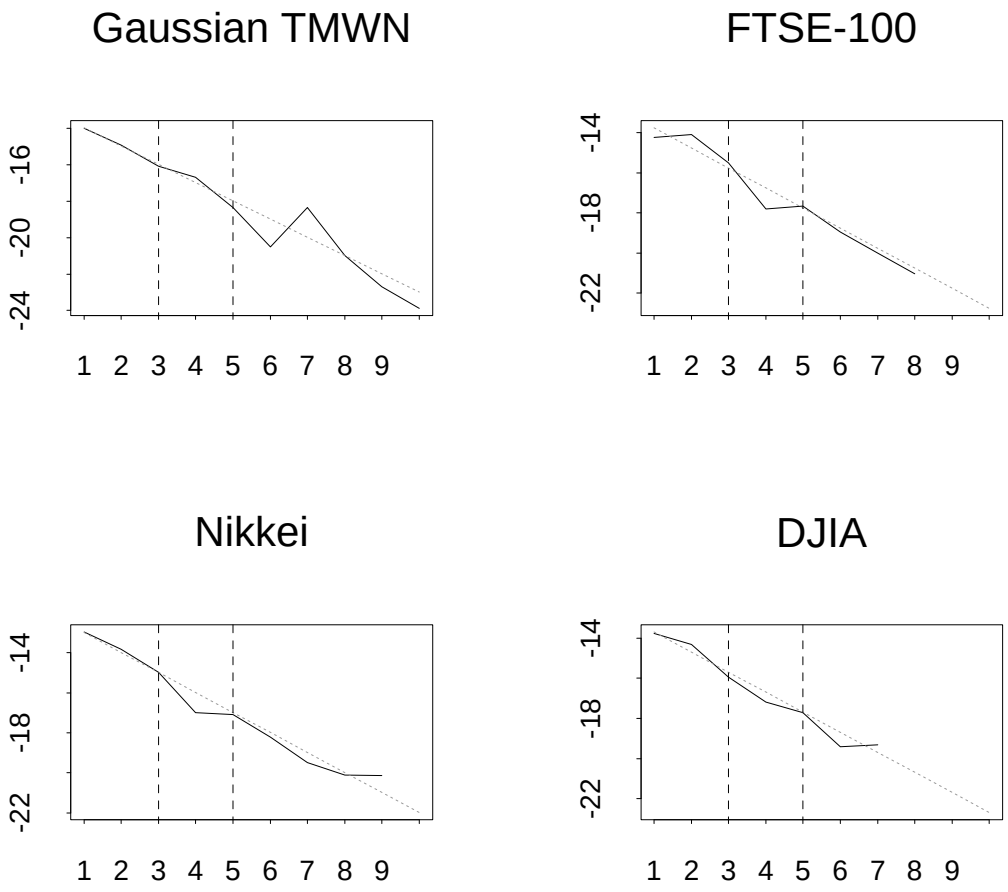

Figure 4: Solid lines: log-scalograms of $X_{t,T}$ (top left), $F_{t,T}$ (top right), $N_{t,T}$ (bottom left) and $D_{t,T}$ (bottom right), plotted against $-j$. Dottes lines: theoretical scalograms if the processes were (time-modulated) white noise (not necessarily Gaussian). Dashed lines: $-j = 3, 5$ (see text for discussion).

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, on average, reasonably close to TMWN. However, FTSE 100, which was "provisionally" modelled as Gaussian TMWN in Section 2, 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 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 (16), a large scalogram at scale $j = -2$ implies a significant contribution of the summand $\bar{S}_{-2}\Psi_{-2}(h)$ 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 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 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 1 shows that the same pattern is present is 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 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.

Full evolutionary Haar spectrum analysis.

Figure 7 shows the estimated evolutionary Haar spectrum of $F'_{t,T}$ = the 2048 last observations of the FTSE 100 index (plotted in Figure 1). 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 $j = -2$, now disappears! The acf of the first 1200 observations of $F'_{t,T}$ is plotted is the left-hand plot of Figure 6. Right-hand plot of Figure 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. The LSW framework provides appropriate tools not only for the local analysis of the log-return data, but also for forecasting. This will be demonstrated in Section 7.

7 Forecasting

7.1 Adaptive forecasting algorithm

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

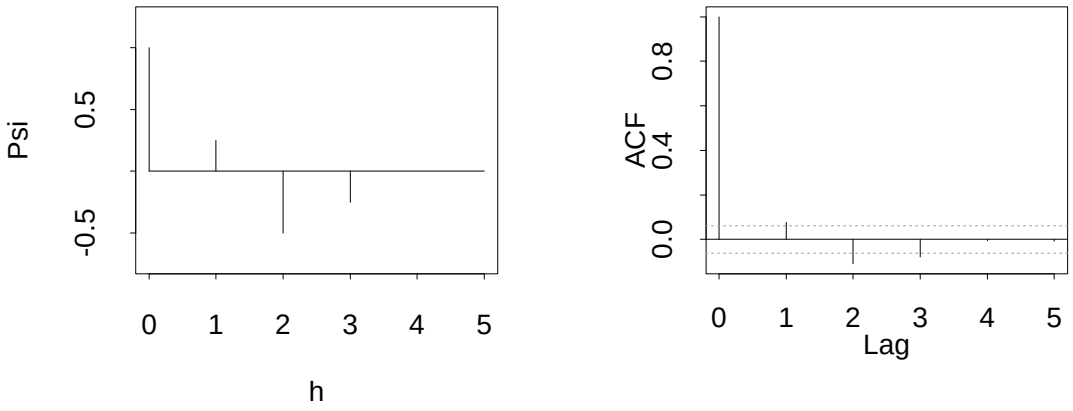

Figure 5: Left-hand plot: $\Psi_{-2}(h)$ for Haar wavelets for $h = 0, 1, \ldots, 5$. Right-hand plot: autocorrelation function for $F_{t,T}$ at lags $0, 1, \ldots, 5$.

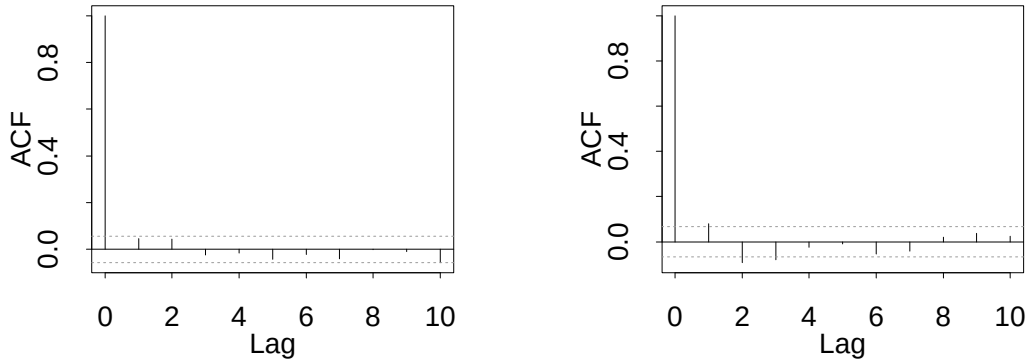

Figure 6: Left-hand plot: sample autocorrelation of $F'_{1,T}, \ldots, F'_{1200,T}$. Right-hand plot: sample autocorrelation of $F'_{1201,T}, \ldots, F'_{2048,T}$.

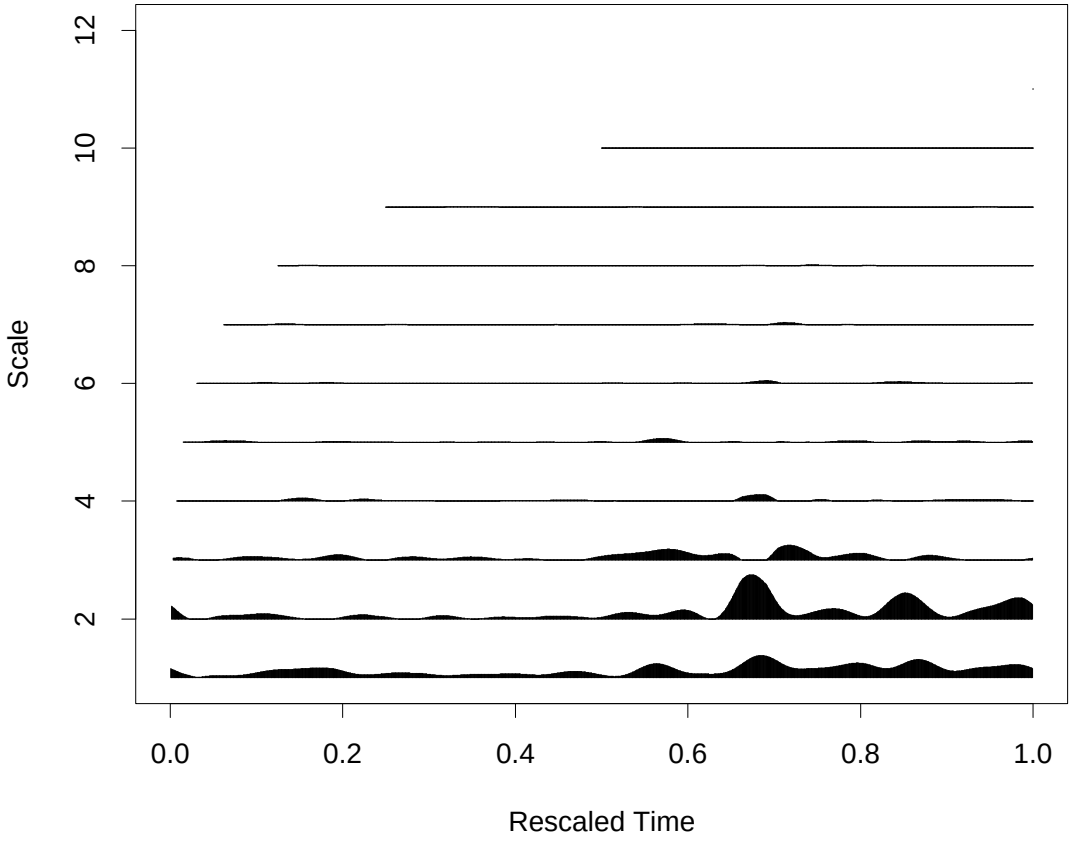

Figure 7: Evolutionary Haar spectrum of $T = 2048$ last observations of FTSE 100 of Figure 1. Smoothing uses TI linear wavelet scheme. X-axis is the rescaled time $z = t/T$, and Y-axis is negative scale $-j = 1, 2, \ldots, 11$.

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 and Higgins (1993)).

A general algorithm for forecasting LSW processes was introduced by Fryźlewicz et al. (2003), and we shall now briefly discuss it in our setting.

Suppose that $X_{t,T}$ is a log-return series which we model as an LSW process and observe up to time t. As the LSW model is linear, it makes perfect sense to consider the n-step-ahead linear predictor

$$\hat{X}_{t+n,T} = \sum_{s=0}^{t} b_s^{(t,n)} X_{s,T}. \tag{28}$$

For simplicity of presentation, we only consider $n = 1$, and denote $b_s^{(t)} := b_s^{(t+n)}$. As in Brockwell and Davis (1987), we find the coefficients $b_s^{(t)}$ by minimising the Mean-Square Prediction Error (MSPE):

$$\mathrm{MSPE}\left(\hat{X}_{t+1,T}, X_{t+1,T}\right) = \mathbb{E}\left(\hat{X}_{t+1,T} - X_{t+1,T}\right)^2.$$

Asymptotically, this minimisation problem is equivalent to solving the system of equations

$$\sum_{m=0}^{t} b_m^{(t)} c\left(\frac{m+n}{2T}, m-n\right) = c\left(\frac{t+n}{2T}, t-n\right), \tag{29}$$

for $n = 0, 1, \ldots, t$ (see Section 4.3 in Fryźlewicz et al. (2003)). Obviously, the local autocovariance c needs to be estimated from the data. This can be done using the algorithm of Section 5, but it seems that we can obtain more accurate forecasts by estimating c using the principle of *adaptive forecasting*: the estimators of Section 5 naturally suffer from edge effects and it is at the right edge where we require the estimates to be particularly accurate in order that the forecasting algorithm perform well.

More precisely, we start with the unsmoothed estimate of the local autocovariance (see Section 5.1 in Fryźlewicz et al. (2003)):

$$\hat{c}(z,\tau) = \sum_{j=-J}^{-1} \left(\sum_i A_{ij}^{-1} I_{i,[zT]}\right) \Psi_j(\tau). \tag{30}$$

We will later smooth this estimate using Gaussian kernel with an appropriately selected bandwidth.

As recalled in Fryźlewicz et al. (2003), Section 5.2, in theory, the best one-step-ahead linear predictor of $X_{t+1,T}$ is given by (28), where $\mathbf{b}_t = \{b_s^{(t)}\}_{s=0,1,\ldots,t}$ solves (29). In practice, we *estimate* each of the prediction coefficients $\mathbf{b}_t$. As we incorporate more and more past observations into the linear predictor (e.g. in (28) we incorporate the whole history of the process), the overall error in estimating the prediction coefficients potentially increases, due to the non-stationarity of the process. On the other hand, the theoretical prediction error $\mathrm{MSPE}(\hat{X}_{t+1,T}, X_{t+1,T})$ decreases. In order to strike a balance between these two types of error, Fryźlewicz et al. (2003) propose to clip the linear predictor at some lag in the past, i.e. to consider

$$\hat{X}_{t+1,T}^{(p)} = \sum_{s=t-p+1}^{t} b_s^{(t)} X_{s,T}. \tag{31}$$

This is reminiscent of the classical idea of AR($p-1$) approximation for stationary processes — here, $p = 1$ loosely corresponds to TMWN, $p = 2$ to time-varying AR(1), and so on (the fact that p corresponds to AR($p-1$) and not to AR(p) is due to the specific form of the autocovariance estimator $\hat{c}$).

Therefore, in order for the forecasting algorithm to work, we have to choose two "nuisance" parameters: lag p and bandwidth h for the local autocovariance smoothing.

The choice is performed using the adaptive forecasting algorithm of Fryźlewicz et al. (2003). Suppose that we observe the series up to time t and want to forecast $X_{t+1,T}$, using an appropriate pair (p, h). We move back by s observations, pretending that only $X_{0,T}, X_{1,T}, \ldots, X_{t-s,T}$ have been observed, and we choose the initial pair of parameters (p_0, h_0). Then, we forecast $X_{t-s+1,T}$ using not only (p_0, h_0), but also the 8 neighbouring pairs of parameters: $(p_0 \pm 1, h_0 \pm \delta), (p_0, h_0 \pm \delta), (p_0 \pm 1, h_0)$, for a fixed value of δ. Since we know the actual value of $X_{t-s+1,T}$, we are able to use a preset criterion to compare the 9 results obtained, and we set (p_1, h_1) to be the pair which gave the best forecast out of the 9. In the next step, we use the pair (p_1, h_1), as well as its 8 neighbours, to forecast $X_{t-s+2,T}$, and then we repeat the update step. In this way, we continue until we reach $X_{t,T}$, when we obtain the pair (p_s, h_s) which we use to perform the actual prediction.

A variety of criteria can be used to compare the performance of the pairs of parameters at each step. Denote by $\hat{X}_{k,T}^{(p)}(h)$ the estimate of $X_{k,T}$ obtained using the pair (p, h). To fine-tune the parameters for the accurate forecasting of the series $X_{t,T}$ itself, a natural choice would be to

choose the pair that minimises

$$d_1(\mathbf{X}, p, h) = \left| X_{k,T} - \hat{X}_{k,T}^{(p)}(h) \right|. \tag{32}$$

However, in order to give preference to forecasts which lie comfortably withing the corresponding prediction intervals, an alternative possibility would be to choose, for example, the pair which minimises

$$d_2(\mathbf{X}, p, h) = \frac{\left| X_{k,T} - \hat{X}_{k,T}^{(p)}(h) \right|}{P_{k,T}^{(p)}(h)}, \tag{33}$$

where $P_{k,T}^{(p)}(h)$ is the length of the corresponding prediction interval. We used d_1 in the simulation study reported below.

Provided that the "training" segment $X_{t-s+1,T}, \ldots, X_{t,T}$ is long enough, (p_s, h_s) should not depend significantly on the initial parameters (p_0, h_0). Fryźlewicz et al. (2003) propose to set s to the length of the largest segment at the end of the series which does not contain any visible breakpoints or "spikes".

The updating step in is accordance with the principle of local stationarity: if a given pair was good at forecasting $X_{k,T}$, we can expect that the same pair, or one of its neighbours, will also perform well in forecasting $X_{k+1,T}$. Also, once the parameters have been fine-tuned on the training set, the forecasting can be performed "online". Indeed, when observation $X_{t+1,T}$ becomes available, we only need to update the pair (p_s, h_s) without having to perform the whole of the "training" step on the past s observations.

The algorithm can be modified by allowing more that one update of parameters at each step. Also, prior knowledge can be incorporated into the model by restricting or penalising certain regions of the parameter space for (p, h).

7.2 Dow Jones example

In this section, we demonstrate the usefulness of the approach by comparing our forecasting methodology to forecasting based on GARCH modelling, on a fragment of the Dow Jones IA series (denoted by $D_{t,T}$ in Section 6 and plotted in Figure 2). This brief simulation study does not aim to show that our approach is superior to GARCH. Instead, we attempt to demonstrate a few interesting features of LSW forecasting.

Suppose that we have already observed 1105 values of the series, and want to perform one-

step-ahead prediction of the series along the segment $D_{1106,T}, \ldots, D_{1205,T}$. In order to do so, we employ the LSW methodology with Haar wavelets. We make an initial guess at the values of p and h: we set $(p, h) = (1, 30)$ (default initial values in our software package, see Section 8 for details on how to obtain the package). Further, we set the criterion function to d_1, 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. As mentioned in Section 7.1, 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 $t = 1105, \ldots, 1204$ 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 routine `garch` from the S-Plus `garch` module with default parameters.

The results of the experiment are presented in Figure 8. The top left plot shows the actual series $D_{1106,T}, \ldots, D_{1205,T}$ (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 model. The bottom right plot in Figure 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.

In the LSW forecasting, the stretches where $p = 1$ wins over $p = 2$ are indicated by one-step-ahead forecasts equal to zero (as in TMWN forecasting). Non-zero forecasts indicate that $p = 2$ is used to perform prediction. The LSW 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

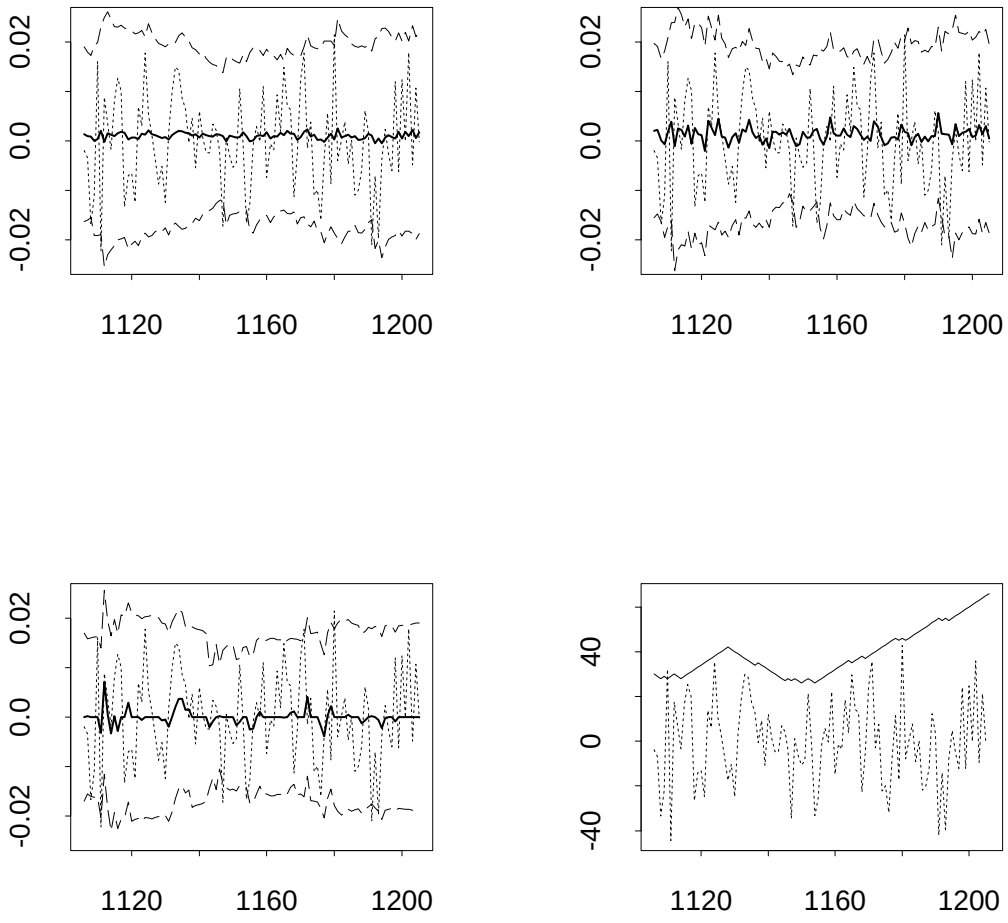

Figure 8: Top left, top right and bottom left: the actual series (dotted line), one-step-ahead forecasts (solid line) and 95% prediction intervals (dashed lines) for AR(1) + GARCH(1,1), AR(16)+GARCH(1,1) and LSW, respectively. Bottom right: actual series $\times 2000$ and the evolution of bandwidth h.

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

Table 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 7.2.

Prediction Errors and the Median Squared Prediction Errors for the three methods are given in Table 2: the LSW method outperforms the other two.

For the LSW method, 92% of observations fall within the corresponding one-step-ahead 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 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.

8 Conclusion

In this article, we have provided theoretical and empirical evidence that stock index returns can be successfully modelled and forecast in the Locally Stationary Wavelet (LSW) framework of Nason et al. (2000). 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 to allow TMWN as a special case of a general LSW process.

We have provided theoretical evidence that the LSW model, being linear and non-stationary, 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 model. We have shown that our new algorithm, specifically designed for financial log-returns, significantly outperforms 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 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 framework using the adaptive forecasting algorithm proposed by Fryźlewicz et al. (2003). 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-based technique has outperformed two GARCH-based methods.

Future ideas. In future research, we intend to examine other distributions of innovations $\xi_{j,k}$ (also combined with "skewed wavelets"), as well as looking at the problem of forecasting volatility in the LSW framework. After completing this work, we were made aware of the recent article by Drees and Starica (2002) in which the authors propose a simple non-stationary model for stock returns which also uses the idea of a time-varying unconditional variance. It would be of interest to investigate the possibility of combining the most attractive features of both models to obtain a further improved linear framework for modelling financial log-returns.

Reproducible research. The S-Plus routines written and used by the author, the data sets analysed in the paper, as well as the contrived standard deviation function of Figure 3 can be downloaded from the associated web page

`http://www.ma.imperial.ac.uk/~pzf/fints/fints.html`

9 Acknowledgements

The author would like to thank Thomas Mikosch, Guy Nason, Rainer von Sachs, Catalin Starica and Sébastien Van Bellegem for very helpful comments and discussions.

A Proofs

Proof of Proposition 3.2.

$$
\begin{aligned}
&|c_T(k/T,\tau) - c(k/T,\tau)| = \\
&\left| \sum_{j=-J}^{-1} \sum_{l} \omega^2_{j,l;T} \psi_{j,l-k}\psi_{j,l-k-\tau} - \sum_{j=-\infty}^{-1} S_j(k/T)\Psi_j(\tau) \right| \leq \\
&\left| \sum_{j=-J}^{-1} \sum_{l} (\omega^2_{j,l;T} - S_j(k/T))\psi_{j,l-k}\psi_{j,l-k-\tau} \right| + \left| \sum_{j=-\infty}^{-J-1} S_j(k/T)\Psi_j(\tau) \right| \leq \\
&\sum_{j=-J}^{-1} \sum_{l} \frac{C_j + L_j(l-k)}{T} |\psi_{j,l-k}\psi_{j,l-k-\tau}| + \sum_{j=-\infty}^{-J-1} S_j(k/T) \leq \\
&\sum_{j=-J}^{-1} \frac{C_j + ML_j 2^{-j}}{T} \sum_{l} |\psi_{j,l-k}\psi_{j,l-k-\tau}| + \sum_{j=-\infty}^{-J-1} S_j(k/T).
\end{aligned}
$$

Now, using assumptions (9) – (12) and Cauchy inequality, we get

$$
\begin{aligned}
&\sum_{j=-J}^{-1} \frac{C_j + ML_j 2^{-j}}{T} \sum_{l} |\psi_{j,l-k}\psi_{j,l-k-\tau}| + \sum_{j=-\infty}^{-J-1} S_j(k/T) = \\
&\quad O\left(\frac{\log(T)}{T}\right) + O\left(\sum_{j=-\infty}^{-J-1} 2^j\right) = O\left(\frac{\log(T)}{T}\right) + O\left(T^{-1}\right) = O\left(\frac{\log(T)}{T}\right).
\end{aligned}
$$

Lemma A.1 *With the assumptions of Lemma 3.1, we have*

$$\sum_{i=-\infty}^{-1} 2^i A_{ij} = 1.$$

Proof. Using Lemma 3.1,

$$\sum_i 2^i A_{ij} = \sum_i 2^i \sum_\tau \Psi_i(\tau)\Psi_j(\tau) = \sum_\tau \delta_0(\tau)\Psi_j(\tau) = 1.$$

Proof of Proposition 4.1. Very similar to the proof of Proposition 4.2 (see below).

Proof of Proposition 4.2. Define $\alpha(z) = \sum_\tau c^2(z,\tau)$. By Assumption (9) and Lemma A.1, we have

$$\alpha(z) = \sum_\tau \sum_{i,i'} S_i(z) S_{i'}(z) \Psi_i(\tau) \Psi_{i'}(\tau) = \sum_{i,i'} S_i(z) S_{i'}(z) A_{i,i'} \leq D \sum_i S_i(z) \leq D^2. \tag{34}$$

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

Proof of Proposition 5.1 We use the orthonormality of $\xi_{j,k}$, the fact that $\mathcal{L}_j \sim M 2^{-j}$, assumptions (9) – (12), and Lemma A.1.

$$\begin{aligned}
&\left| \mathbb{E}\left(\sum_t X_{t,T} \psi_{j,p-t} \right)^2 - \sum_i S_i\left(\frac{p}{T}\right) A_{ij} \right| \leq \\
&\quad \sum_{i=-J}^{-1} \sum_k \left| \omega_{i,k;T}^2 - S_i\left(\frac{p}{T}\right)\right| \left(\sum_t \psi_{i,k-t} \psi_{j,p-t} \right)^2 + \sum_{i=-\infty}^{-J-1} S_i\left(\frac{p}{T}\right) A_{ij} \leq \\
&\quad \sum_{i=-J}^{-1} \frac{L_i M(2^{-i} \vee 2^{-j}) + C_i}{T} A_{ij} + \sup_i \left\{ S_i\left(\frac{p}{T}\right) 2^{-i} \right\} \sum_{i=-\infty}^{-J-1} 2^i A_{ij} \leq \\
&\quad \sum_{i=j+1}^{-1} \frac{L_i M 2^{-j}}{T} A_{ij} + \sum_{i=-J}^{1} \frac{L_i M 2^{-i} + C_i}{T} A_{ij} + D \sum_{i=-\infty}^{-J-1} 2^i A_{ij} \leq \\
&\quad \frac{M 2^{-j}}{T} \left(\sup_{i=-1,-2,\ldots,-J} \left\{ L_i 2^{-i} \right\} \sum_{i=j+1}^{-1} 2^i A_{ij} + \sup_i \left\{ 2^j A_{ij} \right\} \sum_{i=-J}^{1} L_i 2^{-i} + C_i \right) + D \sum_{i=-\infty}^{-J-1} 2^i A_{ij} = \\
&\quad \frac{2^{-j}}{T} \left(O(\log(T)) + O(\log(T)+1) \right) + D \sum_{i=-\infty}^{-J-1} 2^i A_{ij}.
\end{aligned}$$

Let us now turn to the speed of convergence of $\sum_{i=-\infty}^{-J-1} 2^i A_{ij}$.

Haar wavelets. By Theorem 2 of Nason et al. (2000), we have $A_{ij} = 2^i(2^{-2j-1} + 1)$, as $i < j$. Therefore, $\sum_{i=-\infty}^{-J-1} 2^i A_{ij} = O((2^{-j}/T)^2)$.

Other Daubechies' compactly supported wavelets. There is a strong evidence that he above rate is also achieved for other Daubechies' compactly supported wavelets, see Remark 7 in Nason et al. (2000).

Thus, we finally obtain

$$\left| \mathbb{E}\left(\sum_t X_{t,T} \psi_{j,p-t} \right)^2 - \sum_i S_i\left(\frac{p}{T}\right) A_{ij} \right| = O\left(\frac{2^{-j}\log(T)}{T} \right).$$

Now, if $X_{t,T}$ is Gaussian, then

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

as

$$\left| \sum_i S_i(p/T) A_{ij} \right| \leq D$$

by Assumption (9).

Non-Gaussian processes. To extend Proposition 5.1 to non-Gaussian processes, one possibility would be to use Isserlis theorem and the method of cumulants, see Vidakovic (1999), Section 9.4 for details and further references.

References

N. S. Altman. Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, 85: 749–759, 1990.

A. K. Bera and M. L. Higgins. ARCH models: properties, estimation and testing. *J. Economic Surveys*, 7:305–366, 1993.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31: 307–327, 1986.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, 1987.

C. Brooks. Linear and non-linear (non-)forecastability of high-frequency exchange rates. *J. Forecasting*, 16:125–145, 1997.

L. Calvet and A. Fisher. Forecasting multifractal volatility. *J. Econometrics*, 105:27–58, 2001.

D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Nielsen, editors. *Time Series Models in Econometrics, Finance and Other Fields*, volume 65 of *Monographs on Statistics and Applied Probability.* Chapmann & Hall, 1996.

R. Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, 62:139–168, 1996.

I. Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, Pa., 1992.

D. L. Donoho and R. R. Coifman. Translation-invariant de-noising. *Technical Report, Statistics Department, Stanford University*, 1995.

H. Drees and C. Starica. A simple non-stationary model for stock returns. *Submitted*, 2002.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, 2003.

P. Fryźlewicz, S. Van Bellegem, and R. von Sachs. Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Stat. Math*, 55:737–764, 2003.

P. Fryzlewicz and G. P. Nason. Denoising the wavelet periodogram using the Haar-Fisz transform. *Submitted*, 2003.

L. Giraitis, R. Leipus, and D. Surgalis. Recent advances in ARCH modelling. *Submitted*, 2003.

W. Härdle, V. G. Spokoiny, and G. Teyssière. Adaptive estimation for a time inhomogeneous stochastic volatility model. *Submitted*, 2000.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.

W. Kim. Econometric analysis of locally stationary time series models. *Manuscript, Yale University*, 1998.

P. Kokoszka and R. Leipus. Change-point estimation in ARCH models. *Bernoulli*, 6:513–539, 2000.

M. T. Leung, H. Daouk, and A.-S. Chen. Forecasting stock indices: a comparison of classification and level estimation models. *Int. J. Forecasting*, 16:173–190, 2000.

G. S. Maddala and C. R. Rao, editors. *Statistical Methods in Finance*, volume 14 of *Handbook of Statistics*. Elsevier, 1996.

T. Mikosch and C. Starica. Change of structure in financial data, long range dependence and the GARCH modelling. *The Review of Economics and Statistics,* to appear, 2003.

K. G. Murty. *Linear Complementarity, Linear and Nonlinear Programming*. Internet Edition, available at `http://ioe.engin.umich.edu/people/fac/books/murty/linear_complementarity_webbook`, 1988.

G. P. Nason. Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, 58:463–479, 1996.

G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics, vol. 103*, pages 281–300. Springer-Verlag, 1995.

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

H. C. Ombao, J. A. Raz, R. L. Strawderman, and R. von Sachs. A simple generalised cross-validation method of span selection for periodogram smoothing. *Biometrika*, 88:1186–1192, 2001.

S. J. Taylor. *Modelling Financial Time Series*. Wiley, Chichester, 1986.

B. Vidakovic. *Statistical Modeling by Wavelets*. Wiley, New York, 1999.

R. von Sachs and B. MacGibbon. Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scand. J. Statist.*, 27:475–499, 2000.

P. Wasserman. *Advanced Methods in Neural Computing*. Van Nostrand Reinhold, New York, 1993.