Forecasting
Using Locally Stationary Wavelet Processes

Yingfu Xie, Jun Yu and Bo Ranneby

Research Report
Centre of Biostochastics

Swedish University of
Agricultural Sciences

Report 2007:2
ISSN 1651-8543

Forecasting Using Locally Stationary Wavelet Processes

Yingfu Xie[1], Jun Yu and Bo Ranneby

Centre of Biostochastics
Swedish University of Agricultural Sciences
SE-901 83 Umeå, Sweden

Abstract

Locally stationary wavelet (LSW) processes, built on non-decimated wavelets, can be used to analyze and forecast non-stationary time series. They have been proved useful in the analysis of financial data. In this paper we first carry out a sensitivity analysis, then propose some practical guidelines for choosing the wavelet bases for these processes. The existing forecasting algorithm is found vulnerable to outliers, and a new algorithm is proposed to avoid the sensitivity to extreme observations. The new algorithm is shown stable and outperforms the existing algorithm when applied to real financial data. The volatility forecasting ability of LSW modeling based on our new algorithm is then discussed and shown to be competitive with traditional GARCH models.

Keywords: Locally stationary wavelet processes, non-decimated wavelets, sensitivity analysis, GARCH, volatility forecasting.

[1]E-mail address to the correspondence author: yingfu.xie@sekon.slu.se

1 Introduction

Recently, many studies have been published in which time-scale or time-frequency techniques such as wavelet transform have been applied, in addition to traditional time domain analysis of stochastic processes, especially financial time series. One advantage of using wavelet transform is that it depends less on specifications of the dependent structure and distribution of the original series because of the 'whitening' property of wavelets, i.e., wavelet coefficients are often less correlated than the original data, see Vidakovic (1999). Another advantage is that not only stationary time series but also non-stationary ones can be treated in the same framework using wavelets (see e.g. Mallat et al., 1998).

The locally stationary wavelet (LSW) process is a relatively new tool, originally proposed by Nason et al. (2000), which incorporates a class of stochastic processes based on non-decimated wavelets. By defining an evolutionary wavelet spectrum (EWS), which is the analogue of the usual spectrum for stationary processes, the power (variance) of LSW processes can be measured locally over time and scale. Estimation theory has also been developed for the EWS and localized autocovariance. Fryźlewicz (2005) showed that the LSW model can capture most of the stylized facts of financial time series. In addition, Fryźlewicz et al. (2003) developed an algorithm to forecast the LSW processes. The predictor is simply the linear combination of previous observations with the predictor coefficients obtained by minimizing the mean square prediction error (MSPE). Hence the forecasting of non-stationary processes is possible with this algorithm. However, it is found that this algorithm has no protection against the occurrence of outliers, since an inverse of a covariance matrix is unavoidable. The presence of outliers, as a result of the frequently singular matrices, makes it very difficult to evaluate the performance of this algorithm (usually in terms of sample MSPE). In this paper, a new algorithm is proposed to avoid this problem, i.e., we impose suitable restrictions on the predictor coefficients when minimizing the MSPE. Two intuitive restrictions are examined and we focus on the algorithm with the aim of producing a prediction coefficient vector with unit length. Both the new algorithm and the original are tested on real data. The results show that even in one-step-ahead out-of-sample prediction, outliers can appear in the original algorithm, making it difficult to evaluate, while our new algorithm works consistently well.

The volatility forecasting ability of LSW modeling based on our new algorithm is now analyzed. Volatility is a central concept in capital asset pricing, portfolios investment and risk management, especially for products such as options. Volatility forecasting has been the subject of considerable attention

in the last 20 years. In a comprehensive survey summarizing 93 studies, Poon and Granger (2003) classify volatility forecasting models into four categories: HISVOL (Historical volatility), ARCH family (Engle, 1982), ISD (option implied standard deviation) and SV (stochastic volatility). LSW is not included in any of Poon and Granger's categories, so we include a new comparison here. In this paper, we compare the volatility forecasting abilities of the LSW model and GARCH models, including standard GARCH (Bollerslev, 1986), Exponential GARCH (Nelson, 1991) and Regime-Switching GARCH (e.g., Hamilton and Susmel, 1994; Gray, 1996; Xie and Yu, 2005; Xie, 2007), based on log-returns of the S&P500 index. LSW modeling based on our algorithm is shown to be quite competitive with GARCH models.

Our paper is organized as follows: in Section 2 we briefly introduce the definition of the LSW process and the estimation of its EWS. In addition, a sensitivity analysis of the wavelet selection problem is described. The new forecasting algorithm and applications to real financial data are presented in Section 3. The model's ability to forecast volatility is compared with that of the GARCH models in Section 4. Finally, the results and their implications are considered in the Discussion (Section 5).

2 Locally stationary wavelet modeling

In this section, we briefly introduce the definition of LSW processes, EWS and its estimation, mainly based on Nason et al. (2000). The sensitivity of selection of wavelet bases in LSW modeling is examined using numerical examples.

2.1 Locally stationary wavelet processes

Definition 1 (Nason et al. (2000)) *An LSW process is a sequence of doubly-indexed stochastic processes* $\{X_{t,T}\}_{t=0,...,T-1}$ *having the following representation in the mean-square sense*

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

where $\xi_{j,k}$ *is a random orthonormal increment sequence, and* $\psi_{j,k}$ *is a discrete non-decimated family of wavelets for* $j = -1, -2, ..., -J(T), k = 0, ..., T-1$ *based on a mother wavelet* $\psi(t)$ *of compact support. The following properties are also assumed:*

1. *$E\xi_{j,k} = 0$ for all j, k. Hence $EX_{t,T} = 0$ for all t and T.*

2. *$cov(\xi_{j,k}, \xi_{l,m}) = \delta_{jl}\delta_{km}$.*

3. *The amplitudes $\omega_{j,k;T}$ are real constants and for each $j \leq -1$ there exists a Lipschitz-continuous function $W_j(z)$ for $z \in (0,1)$ which satisfies*

$$\sum_{j=-\infty}^{-1} W_j^2(z) < \infty \text{ uniformly in } z \in (0,1)$$

with Lipschitz constants L_j which are uniformly bounded in j and

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

In addition, there exists a sequence of constants C_j fulfilling $\sum_j C_j < \infty$ such that for each T

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

It is well-known that a stationary stochastic process $X_t, t \in \mathbb{Z}$, can be written as

$$X_t = \int_{-\pi}^{\pi} A(\delta) \exp(i\delta t) d\zeta(\delta), \tag{2}$$

where $d\zeta(\delta)$ is an orthonormal increment process (Priestley, 1981). Now, the idea behind the LSW process is to replace the set of harmonics $\{\exp(i\delta t) | \delta \in [-\pi, \pi]\}$ in (2) with a set of non-decimated wavelets $\psi_{j,k}$ and the spectrum $A(\delta)$ by the time-varying $\omega_{j,k;T}$. In fact, LSW processes include all stationary processes with absolutely summable autocovariance as special cases (Nason et al., 2000, Corollary 2). Assumption 3 demands that a smooth $W_j(z)$, as a function of rescaled time, z, controls the variation of $\omega_{j,k}$, as a function of k, so that it cannot change too quickly, allowing effective estimates to be obtained for the model. In this assumption, a rescaled time $z = k/T$ is used which implies that as $T \to \infty$, instead of more and more future data, more detailed local information of $W_j(z)$ are collected. For a more detailed description of the model and of rescaled time, we refer the reader to Nason et al. (2000) and Dahlhaus (1997).

From the definition of LSW processes, direct calculation gives the covariance structure with lag τ as

$$Cov(X_{t,T}, X_{t+\tau,T}) = \sum_j \sum_k \omega_{j,k;T}^2 \psi_{j,k-t} \psi_{j,k-t-\tau}. \tag{3}$$

Considering of $\tau = 0$ leads to Definition 2.

Definition 2 (Nason et al. (2000)) *The Evolutionary Wavelet Spectrum (EWS) of sequence* $\{X_{t,T}\}_{t=0,...,T-1}$ *for infinite sequence* $T \geq 1$ *is defined as*

$$S_j(z) = W_j^2(z), \text{for } j = -1, ..., -J(T), z \in (0, 1).$$

Under Assumption 3 of Definition 1, $S_j(z) = \lim_{T\to\infty} |\omega_{j,[zT];T}|^2$ *and* $\sum_{-\infty}^{-1} S_j(z) < \infty$ *uniformly in* $z \in (0, 1)$.

The EWS measures the local power (variance) at a particular time z and scale j, which is the analogue of usual spectrum for stationary processes. In the stationary case, however, it is independent of time, i.e., $S_j = \omega_{j,k;T}^2 = W_j^2$. The autocovariance is defined as follows:

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

and the local autocovariance with EWS $S_j(z)$ as

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

where the $\Psi_j(\tau) = \sum_{-\infty}^{\infty} \psi_{j,k}\psi_{j,k-\tau}$ is defined as the autocorrelation wavelets. From (3) it can be seen that $\| c_T - c \|_{L_\infty} = O(T^{-1})$ (Nason et al., 2000), which implies that the local autocovariance is the "autocorrelation wavelet" transform of the EWS. We know that for a stationary process, the autocovariance and spectrum are Fourier transforms of each other. This is the analogous result for LSW processes. In particular, the local variance $\sigma^2(z) := c(z, 0) = \sum_{j=-\infty}^{-1} S_j(z)$ since $\Psi_j(0) = 1$ for all values of j.

Definition 3 (Nason et al. (2000)) *The empirical wavelet coefficients of an LSW process* $X_{t,T}$ *are given by*

$$d_{j,k;T} = \sum_{t=0}^{T-1} X_{t,T}\psi_{j,k-t},$$

where $\psi_{j,k}$ *is the same wavelet basis used to build* $X_{t,T}$ *in Definition 1. The wavelet periodogram of* $X_{t,T}$ *is defined as* $I_{j,k} = |d_{j,k;T}|^2$.

The wavelet periodogram is the "building block" for the estimation of EWS. Defining $\mathbf{A}_J$ as the inner product of the autocorrelation wavelet, whose element $A_{j,l} =< \Psi_j, \Psi_l >= \sum_\tau \Psi_j(\tau)\Psi_l(\tau)$, led Nason et al. (2000) to the following proposition:

Proposition 1 *Assuming that the innovations $\xi_{j,k}$ in Definition 1 are Gaussian, we have, for the LSW process $X_{t,T}$,*

$$EI_{j,k} = \sum_l A_{jl} S_l(k/T) + O(T^{-1}). \tag{4}$$

Hence, for the vector of periodogram $\mathbf{I}(k) := \{I_{l,k}\}_{l=-1,\ldots-J}$ and the corrected periodogram vector $\mathbf{L}(k) = \mathbf{A}_J^{-1}\mathbf{I}(k)$,

$$E\mathbf{L}(k) = E\mathbf{A}_J^{-1}\mathbf{I}(k) = \mathbf{S}(k) + O(T^{-1}), \tag{5}$$

where $\mathbf{S}(k) := \{S_j(k/T)\}_{j=-1,\ldots,-J}$. In addition,

$$VarI_{j,k} = 2\{\sum_l A_{jl} S_l(k/T)\}^2 + O(2^{-j}/T). \tag{6}$$

Proposition 1 enables us to use the corrected wavelet periodogram as an unbiased estimate of EWS. However, equation (6) implies that the (corrected) wavelet periodogram is an inconsistent estimate of EWS and needs to be smoothed. Nason et al. (2000) used translation-invariant linear wavelet (TILW) smoothing (Coifman and Donoho, 1995) to do this. Fryźlewicz (2005) compared TILW and cubic B-splines smoothing and showed that they were almost equally powerful. In this paper, we used spline smoothing. For a more detailed description of estimating EWS and local variance, see Fryźlewicz (2005).

2.2 Sensitivity of wavelet selection in LSW modeling

It should be observed that the wavelet periodogram used to estimate the EWS of process $\{X_{t,T}\}$ has to be constructed using the true wavelet basis of $\{X_{t,T}\}$ (Definition 3). In practice this is unrealistic, prompting Nason et al. (2000) to ask the following questions. How can we choose an appropriate wavelet basis on which to build the model, and what happens if we choose an inappropriate basis? Since these questions were posed they have not previously been answered in the literature. In this subsection, we describe a sensitivity analysis, based on numerical examples, that was conducted to demonstrate the effect of selecting the wrong wavelet on the estimate of EWS.

The analysis was conducted as follows. We predetermined a set of wavelets for the comparison and constructed true LSW processes with known EWS and local variance based on these wavelets. For each process, 50 realizations were generated. These wavelets were then applied to all realizations to construct the wavelet periodogram and estimate the EWS. The estimates were then

compared with the true values and the averages (over 50 realizations) of mean square errors (MSE) were summarized.

The wavelets we selected are compactly supported orthogonal wavelets from Daubechies (1992) with different filter lengths, including: Haar, Daublets d4, d8, d12 and d20, Coiflets c6, c12, c18 and c30 and least-asymmetric wavelets s8, s12 and s20. These wavelets are generally representative, with respect to their symmetry and smoothness, of orthogonal wavelets. Descriptions and properties of these wavelets can be found in Daubechies (1992), Bruce and Gao (1996) and Percival and Walden (2000).

Three processes with different wavelets were considered.

- First, a non-stationary process defined by:

$$X_{t,T} = \sum_k \left(\sqrt{S(t/T)} \psi_{-1,k-t} \varepsilon_k \right), \quad T = 2000, \tag{7}$$

  where $S(t/T) = 0.1 + \cos^2(3\pi t/T + 0.25\pi)$, $\psi_{-1,\cdot}$ are the non-decimated wavelet filters at scale $j = -1$ and ε_k is standard Gaussian. By definition, the EWS of this process is simply $S(t/T)$. The results of MSE for this process are presented in Table 1.

- The second process was:

$$X^r_{t,T} = \sum_k \psi_{-r,k-t} \varepsilon_k, \tag{8}$$

  where $\psi_{-r,\cdot}$ are the non-decimated wavelet filters at scale $j = -r$. For these processes $\omega_{j,k;T}$ and W_j equal to 1 when $j = -r$ and otherwise equal to 0. So, the EWS S_j and local variance of $X^r_{t,T}$ are also equal to 1 when $j = -r$ and otherwise equal to 0. In this example, we choose $r = 2$ and $T = 2000$. This sequence is stationary, and the results are presented in Table 2.

- The third process was the concatenation of $X^1_{t,T}$, $X^2_{t,T}$, $X^3_{t,T}$ and $X^4_{t,T}$ as defined in (8), each with length 1024 and standard Gaussian noise. The true EWS should be $S_{-1}(k/T) = 1$ for $k = 1, ..., 1024$ and 0 for other values of k; $S_{-2}(k/T) = 1$ for $k = 1025, ..., 2048$, and so on. The results are shown in Table 3.

From Table 1 it can be seen that the wavelet selection was not so sensitive for the example based on a non-stationary process. In fact, wavelet s8 is quite robust. It has the smallest MSE for eight out of 12 cases (plus another

Table 1: Average (over 50 realizations) MSE ($\times 10^4$) of LSW modeling based on different wavelets. The realizations are generated from the non-stationary process (7) with the same wavelets. The bold numbers are the smallest in their respective columns.

Used\True	Haar	d4	c6	d8	s8	c12	d12	s12	c18	d20	s20	c30
Haar	26.08	34.31	28.72	37.56	37.83	40.36	37.72	42.64	37.86	47.63	41.67	45.82
d4	**24.45**	23.17	**19.44**	24.08	23.77	26.40	22.95	24.89	25.17	29.95	27.94	29.21
c6	27.32	25.45	22.04	26.48	26.60	28.85	25.36	27.04	26.35	32.07	30.27	32.02
d8	30.80	22.47	20.52	**20.73**	20.97	23.09	18.90	20.10	**20.03**	23.20	23.62	24.07
s8	30.49	**22.39**	20.10	20.95	**20.58**	**22.99**	**18.79**	**19.70**	21.28	**22.66**	**23.14**	**23.64**
c12	35.23	25.46	23.22	23.67	23.49	25.76	20.98	22.05	22.98	25.18	25.64	26.36
d12	39.28	27.21	24.74	24.18	23.69	25.84	20.54	21.54	23.73	24.50	24.18	25.21
s12	38.24	26.79	24.20	23.17	22.78	25.31	20.31	21.02	22.89	23.51	24.60	24.76
c18	45.23	31.14	28.56	27.10	26.60	29.47	23.32	24.09	25.88	26.95	28.32	28.20
d20	52.68	36.31	33.80	30.48	32.31	32.32	26.32	27.11	29.18	29.61	30.47	30.39
s20	49.96	33.84	31.39	27.74	27.94	30.59	23.98	24.55	27.47	26.27	28.03	27.87
c30	50.21	34.51	31.58	27.89	28.40	30.66	23.98	24.52	26.47	26.16	27.80	27.33

Table 2: Average (over 50 realizations) MSE ($\times 10^4$) of LSW modeling based on different wavelets. The realizations are generated from the stationary process (8) with the same wavelets. The bold numbers are the smallest in their respective columns.

Used\ True	Haar	d4	c6	d8	s8	c12	d12	s12	c18	d20	s20	c30
Haar	**96.43**	154.89	146.65	251.44	253.72	273.79	339.07	318.89	322.46	394.53	410.97	421.18
d4	106.48	**76.65**	**70.40**	111.61	112.74	123.96	168.26	154.54	155.94	206.42	218.32	222.24
c6	124.14	83.53	78.17	120.94	121.41	132.58	180.32	165.73	164.04	221.91	235.36	236.33
d8	196.11	92.62	84.76	60.25	**62.28**	**63.74**	80.99	71.91	71.79	92.94	105.02	102.97
s8	195.54	91.86	84.21	**59.69**	63.59	64.47	80.81	71.56	71.33	92.22	104.84	103.14
c12	232.57	108.32	99.03	67.21	70.70	67.79	86.02	78.52	76.80	99.68	111.81	109.19
d12	276.36	133.77	125.52	68.064	69.77	67.31	**69.36**	**61.70**	**62.39**	69.26	81.03	76.46
s12	277.19	133.35	124.28	67.28	70.21	66.58	70.41	64.58	62.70	68.66	79.83	75.92
c18	333.59	163.65	151.73	81.11	84.66	78.15	78.36	72.38	72.21	77.82	91.76	83.03
d20	385.80	201.73	189.58	95.56	96.99	89.27	79.81	70.01	70.93	62.82	73.14	65.44
s20	385.28	202.44	190.19	95.21	97.42	90.59	80.23	71.04	72.19	59.79	72.42	64.67
c30	394.37	207.99	195.82	98.04	99.44	91.25	79.92	70.89	71.44	**59.64**	**70.54**	**61.66**

Table 3: Average (over 50 realizations) MSE ($\times 10^4$) of LSW modeling based on different wavelets. The realizations are generated from the concatenated process described in section 2.2, with the same wavelets. The bold numbers are the smallest in their respective columns.

Used\True	Haar	d4	c6	d8	s8	c12	d12	s12	c18	d20	s20	c30
Haar	**123.9**	197.8	184.6	300.3	301.2	290.3	345.5	355.9	357.9	410.4	458.5	392.1
d4	160.3	150.0	**134.3**	180.0	179.9	176.9	202.4	209.7	203.2	253.2	286.6	245.6
c6	185.5	171.4	154.5	205.4	212.1	233.3	228.2	234.6	243.9	290.3	304.2	272.0
d8	232.4	160.2	154.5	**139.5**	159.9	**147.6**	150.6	159.4	157.2	180.1	179.3	177.3
s8	231.3	**155.8**	148.5	143.5	**150.5**	148.2	**146.6**	**158.1**	**155.4**	180.6	**175.7**	**169.9**
d12	307.9	194.4	206.0	167.8	166.7	160.1	167.8	176.0	168.3	178.7	192.0	184.5
s12	307.9	190.5	197.9	165.2	167.2	156.0	158.8	166.5	162.7	**174.1**	183.3	179.9
c18	375.9	238.1	244.9	211.9	213.0	209.2	225.9	219.0	211.9	222.1	247.0	231.5
d20	410.4	275.5	267.7	242.5	241.3	230.3	233.1	229.4	231.9	228.1	258.8	237.6
s20	396.6	266.8	263.8	229.8	233.4	225.9	222.3	220.6	216.4	217.9	241.7	220.4
c30	420.5	315.3	300.5	290.6	291.6	264.8	276.1	266.1	268.3	270.7	291.0	285.5

two cases for d8, which has properties very similar to s8 except symmetry). Wavelets with longer filters are almost all dominated by s8 (except one which is dominated by d8). In addition, the results for the concatenated process (Table 3) exhibit a similar dominance of s8 with respect to averages of MSE (for seven out of 12 cases it is the smallest, plus another two for d8). Note that the concatenated process is also non-stationary.

However, the selection of wavelet bases is more sensitive for the stationary process examined here. For instance, the MSEs increase in the first column of Table 2 nearly monotonically, demonstrating that the more the filter length of a wavelet basis differs from the true one (Haar here) the worse the performance. The first row of MSE in Table 2 shows that using the Haar wavelet in the estimation, the longer filter the true wavelet basis has, the larger MSE the LSW model produces. This even happens with the concatenated process. Other columns and rows exhibit similar trends. In addition, the Daublets, 'd', and Least-asymmetric wavelets, 's', have similar smoothness characteristics, such as vanishing moments (VM), number of derivatives and Hölder exponent (HE), while Coiflet c12 (VM 3, HE 1.45) is close to s8 and d8 (VM 3, HE 1.62 and 1.4), c18 (VM 5, HE 2.21) to s12 and d12 (VM 5, HE 2.19 and 2.12) and c30 (VM 9, HE 3.47) to s20 and d20 (VM 9, HE 3.38 and 3.31) (see Bruce and Gao, 1996, for a description of the Hölder exponent and these properties). Taking the above into consideration, it is always the true wavelet or wavelets with similar smoothness characteristics that produce the best fit. In general, for stationary processes, the selection of wavelet bases is sensitive, and is mainly determined by the smoothness of the wavelet. Care should be taken when choosing the wavelet basis in LSW modeling. In contrast, as can be seen from equation (3), for a wavelet with filter length L the covariance $Cov(X_{t,T}, X_{t+\tau,T}) = 0$ for $\tau > (2^{-J}-1)(L-1)+1$. For a stationary series, this property may be useful when choosing the wavelet filter length, by inspecting the sample covariance and determining the minimum non-trivial scale.

3 Forecasting

In this section we first introduce and assess the original forecasting algorithm for LSW processes published by Fryźlewicz et al. (2003). We found that this algorithm cannot prevent outliers from occurring. This decreases its usefulness, especially when it is evaluated using, e.g., MSPE. Therefore we propose a new algorithm, by imposing restrictions on the predictor coefficients. Here, the original and new algorithms are applied to real financial data and their performance is compared.

3.1 Fryzlewicz's algorithm

Fryźlewicz et al. (2003) developed a forecasting algorithm for LSW processes. Observing that LSW processes have a linear form, a convenient option to consider is a linear predictor for h steps ahead forecast of $X_{t-1+h,T}$, given observations $X_{0,T}, X_{1,T}, ..., X_{t-1,T}$, as

$$\hat{X}_{t-1+h,T} = \sum_{s=0}^{t-1} b_{t-1-s;T} X_{s,T}. \tag{9}$$

The coefficients $b_{j,T}, j = 0, ..., t-1$, are chosen to minimize the MSPE defined as $E(\hat{X}_{t-1+h,T} - X_{t-1+h,T})^2$. That is, the vector $\mathbf{b}_t = (b_{0,T}, ..., b_{t-1,T})'$ ($'$ denoting transposition) is such that

$$\mathbf{b}_t = \arg\min_{\mathbf{b}'_t} \left[(\mathbf{b}'_t, -1) \Sigma_{t+h-1;T} (\mathbf{b}'_t, -1)' \right], \tag{10}$$

where $\Sigma_{t+h-1;T}$ is the covariance matrix of $X_{0,T}, ..., X_{t-1,T}$ and $X_{t-1+h,T}$. Directly taking the derivative over the quadratic form in (10) then equating it to zero leads to a linear equation system for solving $\mathbf{b}_t$

$$\Sigma_{t-1;T} \mathbf{b}_t = \mathbf{C}_{t-1+h} \triangleq \frac{\mathbf{C}_{t-1,h} + \mathbf{C}'_{h,t-1}}{2}, \tag{11}$$

where $\Sigma_{t-1;T}$ is the covariance matrix of $X_{0,T}, ..., X_{t-1,T}$, $\mathbf{C}_{t-1,h}$ is the column vector of covariances between $X_{0,T}, ..., X_{t-1,T}$ and $X_{t-1+h,T}$ and $\mathbf{C}_{h,t-1}$ the vector of covariances between $X_{t-1+h,T}$ and $X_{0,T}, ..., X_{t-1,T}$. These (co)variances can be estimated by estimating the local autocovariance. Fryźlewicz et al. (2003, Remark 8) provided an estimate for these, which was inconsistent and suggested to be smoothed using, for instance, standard kernel smoothing.

In practice, there are two compromises to be made with respect to the above algorithm. First, $\Sigma_{t;T}$ in (10) depends on the amplitudes $\omega_{j,k;T}$, which are not uniquely defined due to the redundancy of non-decimated wavelet families. Based on technical considerations, Fryźlewicz et al. (2003) approximated $(\mathbf{b}'_\mathbf{t}, -1) \Sigma_{t+h-1;T} (\mathbf{b}'_\mathbf{t}, -1)'$ by $(\mathbf{b}'_\mathbf{t}, -1) \mathbf{B}_{t+h-1;T} (\mathbf{b}'_\mathbf{t}, -1)'$, where $\mathbf{B}_{t+h-1;T}$ is a $(t+1) \times (t+1)$ matrix whose (m,n)-th element is

$$\sum_{j=-J}^{-1} S_j(\frac{n+m}{2T}) \Psi_j(n-m)$$

and can be estimated by estimating the EWS S_j. Second, considering the non-stationary nature and local smoothness of the process, it is recommended

that only the most recent p observations in (9) should be used, rather than the entire sequence, i.e.,

$$\hat{X}^{(p)}_{t-1+h,T} = \sum_{s=t-p}^{t-1} b_{t-1-s;T} X_{s,T}. \tag{12}$$

The parameter p, as well as g, the bandwidth of a kernel used to smooth the inconsistent estimator of local autocovariance, can be selected automatically by so-called Adaptive Forecasting (see Fryźlewicz et al., 2003, for detail): Suppose we observe the sequence until $X_{t-1,T}$ and want to predict $X_{t-1+h,T}$. But move first, say, $s+h$ steps backwards and start to predict $X_{t-s,T}$ using $X_{0,T}, ..., X_{t-h-s}$ with initial parameters (p_s, g_s). With some predetermined criterion (usually minimum distance criteria or relative absolute prediction error) and parameter space of (p, g), we obtain the optimal pair of (p_s*, g_s*) and use it as the start value in the next prediction of $X_{t-s+1,T}$, and so on. After this training process, an updated pair (p_1*, g_1*) is finally obtained for the actual forecasting. The number s can be chosen to be the length of the largest segment at the end of sequence containing no apparent visually observable breakpoints. When feasible, we can run the algorithm several times using (p_1*, g_1*) as the start value in the next iteration until it performs reasonably well.

3.2 The new algorithm

Fryźlewicz's algorithm may work well for short forecasting horizons (usually small p's) and carefully chosen parameters (see Fryźlewicz et al., 2003, and Fryźlewicz, 2005, for examples). However, we find that in this algorithm extraordinary high value of $\mathbf{b}_t$ is often obtained when solving (11) because the covariance matrix often become singular, even for moderately large value t in (9) (or p in (12)). A similar problem is well known in linear regression when there are many regressors. Consequently, the forecasts predict abnormally large values (outliers). After some investigation we found that this problem was difficult to circumvent without artificially intervening in a case-by-case manner. There are a number of potential remedies for this, for example, avoiding the use of the quadratic form of prediction error during the minimization, and instead selecting an alternative yet robust criterion. Further, it is also possible to avoid the use of MSPE when evaluating the performance of this forecasting algorithm, thus reducing the effect of outliers. However, it may be of more interest to prevent outliers from occurring in the first place. Here a new forecasting algorithm is proposed to achieve this.

We suggest imposing some restriction on the predictor coefficients $\mathbf{b}_t$ when minimizing the quadratic form of (10). Particular forms of restrictions should be specified for different problems to obtain a solution. An obvious constraint in this is to require the sum of $\mathbf{b}_t$ to be one, or

$$\mathbf{b}_t'\mathbf{1} = 1, \tag{13}$$

where $\mathbf{1}$ is the unit vector with same length as $\mathbf{b}_t$. This actually works as a weighted average predictor with data-driven coefficients. The solution of (10) with constraint (13) is quite simple using the Lagrangian Multiplier (LM) method. $\mathbf{b}_t$ is the solution of the following equation system

$$\begin{pmatrix} \Sigma_{t-1;T} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{b}_t \\ -\frac{1}{2}\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{t-1+h} \\ 1 \end{pmatrix}, \tag{14}$$

where λ is the Lagrangian multiplier. However, imposing constraint (13) cannot prevent the excessively predictor coefficients from occurring. Hence, it does not fit our purpose and so we will mainly focus on another convenient choice, namely the requirement for a unit length of vector $\mathbf{b}_t$, i.e.,

$$\mathbf{b}_t'\mathbf{b}_t = 1. \tag{15}$$

It should be mentioned that, by imposing a restriction to $\mathbf{b}_t$, the parameter space of $\mathbf{b}_t$ is reduced and we may obtain only local maxima. In addition, the solving of $\mathbf{b}_t$ under condition (36) is more complicated. Similarly using the LM method, we obtain

$$\mathbf{b}_t = (\Sigma_{t-1;T} - \lambda\mathbf{I})^{-1}\,\mathbf{C}_{t-1+h}, \tag{16}$$

where $\mathbf{I}$ is the identity matrix of same size as Σ, λ is again the Lagrangian multiplier and satisfies

$$\mathbf{C}_{t-1+h}'(\Sigma_{t-1;T} - \lambda\mathbf{I})^{-1}(\Sigma_{t-1;T} - \lambda\mathbf{I})^{-1}\mathbf{C}_{t-1+h} = 1 \tag{17}$$

and

$$\Sigma_{t-1;T} - \lambda\mathbf{I} > 0 \text{ (positive definite)}. \tag{18}$$

It is difficult, if not impossible at all, to obtain an analytic solution of λ from equation (17). Instead we use numerical computing. The numerical experiments show that in general this equation is very unsmooth around zero and there are usually multiple roots near zero for λ depending on the covariance matrix. Alternatively, we can try to solve this by minimizing the

Table 4: MSPE ($\times 10^5$) of the two algorithms with different wavelets bases in one-step-ahead forecasting of FTSE 100 data $F_{1659}, ..., F_{1758}$.

	Haar	d4	c6	s8	d10
Fryźlewicz's algorithm	168.17	43.66	753.84	16.78	24.27
Unit-length algorithm	26.05	23.79	27.07	24.55	24.01

function $f(\lambda) \triangleq (\mathbf{C}'_{t-1+h}(\Sigma_{t-1;T} - \lambda\mathbf{I})^{-1}(\Sigma_{t-1;T} - \lambda\mathbf{I})^{-1}\mathbf{C}_{t-1+h} - 1)^2$. In practice, selecting one of the minima of $f(\lambda)$ over the interval $(-1, 1)$ satisfying (18) gives quite satisfactory result in our forecasting applications, using the optimization routine in R (R Development Core Team, 2005) or S-plus®. The two practical settings, approximating the MSPE by $(\mathbf{b}'_\mathbf{t}, -1)\mathbf{B}_{t+h-1;T}(\mathbf{b}'_\mathbf{t}, -1)'$ and using the 'clipped' series (12) for prediction, are also used in our procedure. The adaptive forecasting procedure to obtain the nuisance parameters is also applied.

3.3 Applications and comparison of the algorithms

In this section of study, we compare these algorithms and demonstrate the usefulness of the new algorithm and the need to avoid the influence of outliers in the evaluation.

In the first experiment, our algorithm with $\mathbf{b}_t$ obtained from (16), denoted as the unit-length algorithm, and Fryźlewicz's algorithm (11) are applied to the log-returns of daily FTSE 100 index from 22/23 October 1992 to 10/11 May 2001, F_t. We start from F_{1658} and forecast one step ahead for every step, resulting in a total of 100 steps. (This sequence was also investigated by Fryźlewicz, 2005). The parameters are automatically chosen by the adaptive forecasting procedure for given starting setups, for example, start values of p and g and so on. The MSPE is presented in Table 4. From the sensitivity analysis, we may use wavelet s8 as an example. Wavelets Haar, d4, c6 and d10 are also included here for comparison.

Table 5 shows the results for SP_t, the S&P500 index daily log-returns from 2 Jan 1990 to 29 Dec 2000, from the Center for Research in Securities Prices (CRSP) database. The one-step-ahead forecasting starts from a half sample SP_{1390} and runs for 100 steps. Another example with a longer forecasting horizon is also considered. Table 6 shows the MSPE of these two algorithms for a ten-step-ahead forecast of the same S&P500 return series.

Clearly, the performance of the original algorithm of Fryźlewicz et al. (2003) is severely affected by the outliers. Even in the one-step-ahead fore-

Table 5: MSPE ($\times 10^6$) of the two algorithms with different wavelets bases in one-step-ahead forecasting of S&P500 data $SP_{1391}, ..., SP_{1490}$

	Haar	d4	c6	s8	d10
Fryźlewicz's algorithm	1241.32	44.83	46.52	47.38	44.16
Unit-length algorithm	47.75	46.25	40.84	45.30	48.81

Table 6: MSPE ($\times 10^6$) of the two algorithms with different wavelets bases in ten-step-ahead forecasting of S&P500 data $SP_{1400}, ..., SP_{1499}$.

	Haar	d4	c6	s8	d10
Fryźlewicz's algorithm	215.81	54.82	202.15	581.09	101.06
Unit-length algorithm	38.47	40.58	34.41	43.80	49.65

casts, where smaller p's are usually adapted, the algorithm still causes this problem. For example, a single outlier (0.8442) dominates near 85% MSPE in the FTSE experiment with the c6 wavelet basis. The outlier (-0.3404) gives a 96.36% MSPE in the S&P500 experiment based on the Haar wavelet. Moreover, most results in the ten-step-ahead experiment (Table 6) for Fryźlewicz's algorithm are unsatisfactory due to outliers. Once again, such outliers are difficult to predict and avoid without artificially intervening on a case-by-case basis. Note that in the real calculation Fryźlewicz et al. (2003) deliberately set their forecasts to zeros when the parameter p, the number of observations to be used in (12), obtained is less than or equal to the corresponding forecasting horizon h. We do not see the rationale behind this and do not follow this practice. For instance, in the one-step-ahead cases, whenever $p = 1$, Fryźlewicz et al. (2003) forecast zero values, while the forecasts for our algorithm are the previous observations from (15). Otherwise, the setup for the two algorithms was identical for each experiment. For different setups, the results of Fryźlewicz's algorithm can change dramatically for different wavelets, depending on the occurrences of outliers, while results for the unit-length algorithm vary only slightly. However, the overall performance of Fryźlewicz's algorithm is similar to (if not worse than) that reported here. As mentioned previously, the outlier problem persists in the 'average algorithm' with constraint (13) and its performance is not significantly better than Fryźlewicz's algorithm. We, therefore, did not include it in this study. From these experiments the unit-length algorithm shows promise; it works consistently and outperforms Fryźlewicz's algorithm in most cases.

4 Application to volatility forecast

Our aim in this section is to illustrate the volatility forecasting ability of LSW modeling. After obtaining the forecasts described in the previous section, the next step is to obtain volatility forecasts by estimating the EWS of the sequence together with the forecasts. Fryźlewicz's algorithm is not suitable for this due to the occurrence of outliers. This problem becomes more apparent in volatility forecasts since through wavelet transform more of them will be affected by a single outlier derived from the sequence forecasting. The performance of Fryźlewicz's algorithm can be very poor in terms of MSPE. Therefore, only the unit-length algorithm was used in this study. From the sensitivity analysis presented in Section 2.2, wavelet s8 is chosen as a representative and some other wavelets are also discussed.

Here, GARCH models are compared to our LSW model. Besides the standard GARCH(1,1) model, GARCH(1,1) with student-t as a conditional distribution (GARCH-t(1,1)) and exponential GARCH (EGARCH(1,1)) model are also considered. The GARCH-t model may capture the occurrence of fat tails that are often observed in financial sequences. For the EGARCH model, the process may be written as:

$$r_t = \sigma_t \eta_t, \tag{19}$$

$$\log(\sigma_t^2) = \alpha_0 + \sum_{i=1}^{p} \alpha_i \frac{|r_{t-i}| + \gamma_i r_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{q} \beta_j \log(\sigma_{t-j}^2), \tag{20}$$

where, σ_t is the conditional variance of r_t conditioning on information up to time $t-1$, η_t is usually assumed to be standard Gaussian and $\alpha_i, i = 0, 1, ..., p$, and $\beta_j, j = 1, ..., q$, are parameters to be estimated. From equation (20), not only is the positive parameter constraint of the GARCH model unnecessary, the leverage effect exerted by bad news (negative shocks) tending to have a greater impact on the volatility than good news (positive shocks), is also incorporated, by introducing (usually negative) parameters γ_j. See Nelson (1991) for detail. Finally, we will consider the Regime-Switching GARCH (RS-GARCH) model too, which is defined by:

$$r_t = \sigma_t \eta_t,$$

$$\sigma_t^2 = \alpha_0(Y_t) + \sum_{i=1}^{q} \alpha_i(Y_t) r_{t-i}^2 + \sum_{j=1}^{p} \beta_j(Y_t) \sigma_{t-j}^2. \tag{21}$$

In (21), a regular (stationary, ergodic), finite state (assuming two states in our experiment) Markov chain $\{Y_t\}$ is incorporated into the conditional variance equation. This model can remedy the volatility persistence problem of

the GARCH model and has been shown to be superior to GARCH in model explanation and volatility forecasting in some case (e.g., Klaassen, 2002).

We use the module FinMetrics in S-plus® to generate estimates and forecasts for the ordinary GARCH models. With respect to the RS-GARCH model, based on the suggestions of Klaassen (2002) and Xie and Yu (2005), suppose we want to predict the volatility at time t based on information up to $t-1$, I_{t-1}. By letting $\lambda_{it} = Pr(Y_t = i|I_{t-1})$ and making use of the Bayesian rule, it is relatively straight-forward to obtain the predicted regime

$$\lambda_{it} = \sum_{j=1}^{d} p(j,i) \frac{f_j(r_1, ..., r_{t-1})\lambda_{jt-1}}{\sum_{k=1}^{d} f_k(r_1, ..., r_{t-1})\lambda_{kt-1}}, \tag{22}$$

where $p(j,i)$ is the Markov transition probability, d is the number of states (regimes), $f_j(r_1, ..., r_t)$ is the density (assumed to be Gaussian when MLE is calculated) of r_t given all previous observations $r_1, ..., r_{t-1}$ and regime j. This procedure continues recursively. After obtaining the regime forecast, the volatility forecasting of the RS-GARCH model is similar to that of the GARCH model. Further details of the RS-GARCH model are presented in Gray (1996), Francq et al. (2001), Xie and Yu (2005), and Xie (2007).

The data are the same S&P500 return series that we considered in Section 3.2, with a total of 2780 samples. To perform the out-of-sample forecast, starting from a half sample ($t = 1390$, 29 June 1995), we estimate the parameters using all previous observations and forecast 1 to 50 steps ahead using all these models. After every 50 steps, we update the data, re-estimate the parameters and forecast again. The sample MSPEs with respect to the true volatilities ($\sigma_t^{2,*}$, see equation (23) below) for all forecasting horizons were summarized as an evaluation criterion.

The definition of true volatility is a very important issue in volatility forecasting. The standard way is to use the square of returns (or returns minus the sample mean if it is not zero) as an approximation (see, *inter alia*, Poon and Granger, 2003, and Gokcan, 2000). Andersen and Bollerslev (1998), however, argued that using this definition generally leads to a model with poor goodness-of-fit, because this measurement typically displayed a large degree of idiosyncratic, observation-by-observation variation (see particularly their Figure 1). They instead proposed the use of cumulative squared intraday returns and showed strikingly improvements. Along the same line McMillan and Speight (2004) reassessed the performance of GARCH models and argued that with this measure of true volatility GARCH models outperformed smoothing and moving average models in a data set of 17 daily exchange rate series. However, the high frequency data are not always available to use

this technique. Alternatively, Stărică (2003) and Stărică and Granger (2005) used grouped realized volatility, $\sum_{i=1}^{h} r_{t+i}^2$ as the true volatility for forecasts of $(\sigma_{t+1}^2 + ... + \sigma_{t+h}^2)$, in the hope that the averaging (summation) would cancel out some of the idiosyncratic noise in the daily squared returns. It is worthy noting that Stărică and Granger (2005) gave up the stationarity assumption of whole S&P500 returns (from 1928 to 2000) and argued for the use of stationary, or even independent, identically distributed (i.i.d) sequences to model the return series piece by piece. The i.i.d sequences have a mean of zero and constant variances.

Recall that the S&P500 return series is non-stationary, but with a locally stationary structure. We propose that it is perhaps more appropriate to define the true volatility as a local mean of squared observations over a symmetric interval $(t-m, t+m)$ around the observation at time t for some positive integer m, i.e.,

$$\sigma_t^{2,*} = \frac{1}{2m+1} \sum_{i=-m}^{m} r_{t+i}^2 - (\frac{\sum_{i=-m}^{m} r_{t+i}}{2m+1})^2. \tag{23}$$

This volatility definition can also be applied to stationary processes. However, it is particularly suitable for processes with a locally stationary structure, smooth evolution of the variance or even variances which have a linear trend. Let $l = 2m + 1$ be the interval length. A fairly large l should be used for stationary processes, and a smaller one for non-stationary processes. In our application, $l = 5, 11, 19$ and 31 were used, and their differences are discussed below.

The ratios of the MSPE of different GARCH models, over that of the LSW modeling with the s8 wavelet, are presented in Figure 1. The ratio for the RS-GARCH model is not included in the figure to facilitate interpretation of the graph. The ratios are usually over five for most forecasting horizons. Perhaps surprisingly, in-sample estimation of the data shows only one regime is visible. Regime prediction also always adheres to a single regime. In this case, the model is too complicated to produce an accurate forecast. Clearly from Figure 1, the volatility forecasting for the LSW model is promising. For wavelet s8 and $l = 5$, it gives more accurate forecasts than GARCH(1,1) for most of the forecast horizons. It is more than twice as good around $h = 22$. It is also competitive for small h's, where it is well-known that GARCH can give good forecasts. The GARCH model only outperforms it for long horizons and large l. Recall that during the adaptive forecasting to obtain the nuisance parameters p in (10) and bandwidth g for kernel smoothing, we move h steps backwards. In a sense, we use parameters obtained from quite 'old' information for the real forecasting when h is large. This seems to be justifiable since non-

stationary processes are unsuitable for long-term forecasts.

For different lengths of the intervals defining the volatility, the comparison differs slightly. With respect to the size of the ratios, the performance of GARCH(1,1) is, to some extent, improved with a 'smoother' volatility definition (large l), while LSW modeling favors a 'sharp' definition, due to the fact that wavelet methods are capable of preserving more detailed information about processes. The GARCH-t model behaves in a similar way to the standard GARCH model, but is usually less accurate. In contrast, the EGARCH(1,1) model generally performs better than GARCH(1,1) except for $l = 31$. With respect to other wavelets, similar figures for c6, s12, d12 and c30 are also obtained (not reported here for sake of space). In general, short wavelet (in the sense of filter length, e.g. c6, s8) preserves details of information and performs better with 'sharp' true volatility, while long wavelets (like s12) performs relatively better with 'smooth' volatility. The result for wavelet d12 was not so good, perhaps because it is a little too long and we used the symmetric volatility definition in (23) although wavelet d12 is very asymmetric. Over all, it is clear that wavelet c30 is not recommended for LSW forecasting for S&P500 data. It is too 'long' for a non-stationary process. Once again, wavelet s8, which is smooth, near-symmetric and has a support of medium length, tends to be a good choice.

5 Discussions

As Nason et al. (2000) and Fryźlewicz (2005) pointed out, LSW modeling was developed not because it is superior to other approaches, but as an attractive alternative, because of properties such as its linearity, local stationary nature and the availability of estimation and forecasting methods. It is particularly useful when time and scale must be jointly considered and/or local patterns are of interest.

However, the first problem with LSW modeling for real data is choosing a suitable wavelet basis. Theoretical results concerning the estimation of EWS and local covariance (Nason et al., 2000) are all based on the assumption that we know the true wavelet basis. It is unclear what will happen if a wrong wavelet is used. We conducted a sensitivity analysis of the selection of different wavelet bases, based on simulated data. The criterion was the mean of MSE (over 50 realizations) for estimating the EWS. For non-stationary processes, the selection of wavelets was found to be not sensitive, and wavelet s8 outperformed the others in most of cases. Together with its outstanding performance in the volatility forecasting of S&P500 returns in Section 4, s8

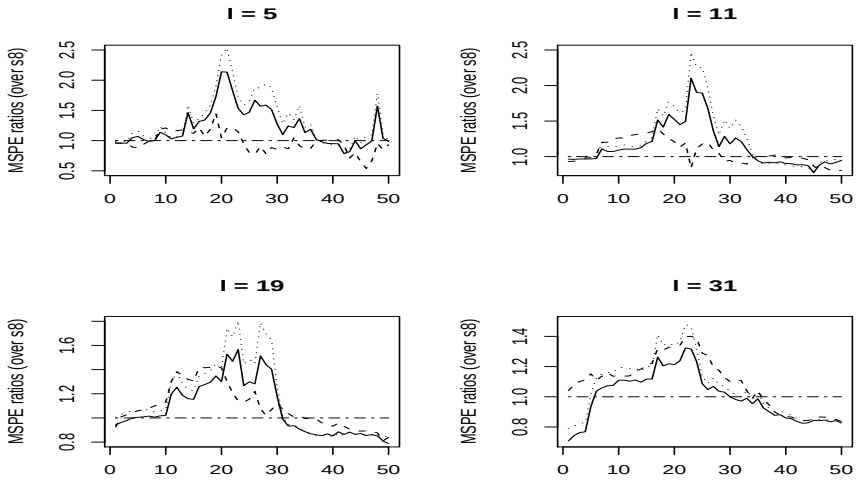

Figure 1: Ratios of GARCH(1,1) (solid), EGARCH(1,1) (dashed) and GARCH-t(1,1)(dotted) MSPEs divided by corresponding MSPE from the LSW modeling with s8 wavelet basis in forecasting S&P500 data, and unit line (dash-dotted), against the forecasting horizons. The lengths of intervals l, defining the true volatility (23), are 5, 11, 19, and 31, respectively.

could be a good candidate for any non-stationary process with respect to LSW modeling. For the stationary process, the wavelet selection is more sensitive. There is no dominating wavelet family and the true wavelet or a wavelet with a similar smoothness characteristic always provides the best fit. However, we observed a 'cutting' property which indicates that for a wavelet whose filter length is L, the covariance $Cov(X_{t,T}, X_{t+\tau,T}) = 0$ for $\tau > (2^{-J} - 1)(L-1)+1$. For a stationary series, it may help to identify the right wavelet filter length by inspecting the sample autocorrelation and determining the minimum non-trivial scale.

We propose a new forecasting procedure with LSW processes in order to avoid the outliers that are produced by the original algorithm developed by Fryźlewicz et al. (2003). We suggest that some restrictions should be imposed on the choice of predictor coefficients $\mathbf{b}_t$ in (9) or (12) when minimizing the MSPE, while focusing on the unit-length algorithm with constraint $\mathbf{b}_t'\mathbf{b}_t = 1$. Applications to real data show that outliers in Fryźlewicz's algorithm make the evaluation of this algorithm very difficult. In contrast, the unit-length algorithm works consistently and outperforms the original in most of cases. Of course, constraints other than (13) or (15) are other possible considerations.

Note also that instead of imposing constraints on $\mathbf{b}_t$ to prevent outliers occurring, one can also try to use more robust minimizing and evaluation criteria instead of the quadratic form.

Based on our unit-length algorithm, the volatility forecasting ability of LSW modeling was investigated and compared with that of the GARCH models. For the example of S&P500 return data, we introduced a new definition of true volatility (23) by realizing the possible non-stationary nature of this sequence. Comparison using MSPE shows that the volatility forecasting with the LSW model is promising. With wavelet s8 (among others), it outperforms GARCH models for a large range of forecasting horizons. In general, a fair comparison is not easy to achieve because even for the same pair of models, their relative merits can vary with differences in data frequency, data size, forecast horizon, true volatility definition, evaluation criterion and other factors. The overall ranking from Poon and Granger (2003) suggests that ISD performs best, followed by HISVOL and GARCH models (which perform roughly equally well, although other studies cited in their review come to different conclusions regarding GARCH and HISVOL models). ISD entails option prices and is not generally available for other assets. From our experiment, it seems that LSW forecasting can be a valuable alternative, especially in a non-stationary or locally stationary situation. Note that other methods for non-stationary processes have also been developed. For instance, Stărică and Granger (2005) divided a long, non-stationary process into homogeneity intervals, to each of which a stationary, even i.i.d sequence was modeled. They also showed the superiority of forecasting based on this methodology over the GARCH (1, 1) model. Fryźlewicz et al. (2006) used (*inter alia*) wavelet shrinkage to study processes with stepwise variance. We conjecture that such a technique could also be applied to stochastic processes with smoothly evolving variances, such as LSW processes, and suggest that this possibility should be explored in future work.

Acknowledgements

The authors thank the High-Performance Computing Center North (HPC2N) at Umeå University, Sweden for providing computational assistance and Dr. Magnus Ekström for help with this. The computation in this paper is benefited from the open S-plus® code (for the Haar wavelet only) accompanying the paper by Fryźlewicz et al. (2003). The codes for other wavelets are available from the authors upon request.

References

Andersen, T.G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *Internat. Econo. Review* **39**:4, 885-905.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. (2001). The distribution of realized stock return volatility. *J. Financial Econ.* **61**, 43-76.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**, 307-328.

Bruce, A. and Gao, H.-Y. (1996). *Applied Wavelet Analysis With S-Plus*. Sringer-Verlag, New York.

Coifman, R.R. and Doholo, D.L. (1995). Translation-invariant de-noising. *Lect. Notes Statist.* **103**, 125-150.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25**, 1-37.

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

Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**:4, 987-1007.

Francq, C., Roussignol, M. and Zakoian, J. (2001). Conditional heteroskedasticity driven by hidden Markov chains. *J. Time Series Anal.* **22**, 197-220.

Fryźlewicz, P. (2005). Modeling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Statist.* **32**, 503-528.

Fryźlewicz, P., Sapatinas, T. and Subba Rao S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika* **93**:3, 687-704.

Fryźlewicz, P., Van Bellegem, S. and von Sachs, R. (2003). Forecasting nonstationary time series by wavelet process modeling. *Ann. Inst. Statist. Math.* **55**, 737-764.

Gray, S.F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *J. Financial Econ.* **42**, 27-62.

Gokcan, S. (2000). Forecasting volatility of emerging stock markets: Linear versus non-linear GARCH models. *J. Forecasting* **19**, 499-504.

Hamilton, J.D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regimes. *J. Econometrics* **64**, 307-333.

Heynen, R.C. and Kat, H. M. (1994). Volatility prediction: A comparison of stochastic volatility, GARCH(1,1) and EGARCH(1,1) models. *J. Derivatives*, 50-65.

Klaassen, F. (2002). Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Econ.* **27**, 363-394.

Mallat, S.G., Papanicolaou, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Statist.* **26**, 1-47.

McMillan, D.G. and Speight, A.E.H. (2004). Daily volatility forecasts: Reassessing the performance of GARCH models. *J. Forecasting* **23**, 449-460.

Nason, G.P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Statist. Soc. B* **62**, 271-292.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**:2, 347-370.

Percival, D.B.and Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge.

Poon, S.-H. and Granger, C.W.J. (2003). Forecasting volatility in financial markets: A review. *J. Econ. Literat.* **XLI**, 478-539.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In: D.R. Cox, O.E. Bandorff-Nielsen and D.V. Hinkley (eds.) *Statistical Models in Econometrics, Finance, and Other Fields*. Chapman and Hall, London, pp. 1-67.

Stărică, C. (2003). Is GARCH(1,1) as good a model as the Nobel prize accolades would imply? Manuscript.

Stărică, C. and Granger, C.W.J. (2005). Nonstationarities in stock returns. *Reviews Econ. Statist.* **87**:3, 503-522.

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

Xie, Y. (2007). Consistency of maximum likelihood estimators for the regime-switching GARCH models. To appear in *Statistics*.

Xie, Y. and Yu, J. (2005). Consistency of maximum likelihood estimators for the reduced regime-switching GARCH models. Research report 2005:2, Centre of Biostochastics, Swedish University of Agricultural Sciences.