Biometrika (2006), **93,** 3, *pp.* 687–704

Printed in Great Britain

A Haar–Fisz technique for locally stationary volatility estimation

By PIOTR FRYZLEWICZ
Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.
p.z.fryzlewicz@bris.ac.uk

THEOFANIS SAPATINAS
Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, CY 1678 Nicosia, Cyprus
t.sapatinas@ucy.ac.cy

and SUHASINI SUBBA RAO
Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.
s.subbarao@bris.ac.uk

Summary

We consider a locally stationary model for financial log-returns whereby the returns are independent and the volatility is a piecewise-constant function with jumps of an unknown number and locations, defined on a compact interval to enable a meaningful estimation theory. We demonstrate that the model explains well the common characteristics of log-returns. We propose a new wavelet thresholding algorithm for volatility estimation in this model, in which Haar wavelets are combined with the variance-stabilising Fisz transform. The resulting volatility estimator is mean-square consistent with a near-parametric rate, does not require any pre-estimates, is rapidly computable and is easily implemented. We also discuss important variations on the choice of estimation parameters. We show that our approach both gives a very good fit to selected currency exchange datasets, and achieves accurate long- and short-term volatility forecasts in comparison to the GARCH(1, 1) and moving window techniques.

Some key words: GARCH model; Haar wavelet; Locally stationary model; Variance-stabilising transform; Wavelet thresholding.

1. Introduction

Log-returns on speculative prices, such as stock indices, currency exchange rates, share prices, and so on, often exhibit the following well-known properties: the sample mean of the observed series is close to zero; the marginal distribution is roughly symmetric or slightly skewed, has a peak at zero, and is heavy tailed; the sample autocorrelations are ‘small’ at almost all lags, although the sample autocorrelations of the absolute values and squares are significant for a large number of lags; and volatility is ‘clustered’, in that days of either large or small movements are likely to be followed by days with similar characteristics.

To capture the above properties, one needs to look beyond the stationary linear framework, and in order to preserve stationarity a large number of nonlinear models have been proposed. Among them, two branches are by far the most popular; namely the families of ARCH (Engle, 1982) and GARCH (Bollerslev, 1986; Taylor, 1986) models, as well as the family of 'stochastic volatility' models (Taylor, 1986). For a review of recent advances in ARCH, GARCH and stochastic volatility modelling, we refer the reader to Fan & Yao (2003) and Giraitis et al. (2005).

Although stationarity is an attractive assumption from the estimation point of view, some authors point out that the above properties can be better explained by resorting to nonstationary nonlinear models; see for example Kokoszka & Leipus (2000), Mikosch & Stărică (2004) and Dahlhaus & Subba Rao (2006). Underlying all these approaches is the observation that, given the changing pace of the world economy, it is unlikely that log-return series should stay stationary over long time intervals.

However, an interesting question which arises once one relaxes the assumption of stationarity is whether nonlinearity is still needed to model log-returns acccurately, or whether it is sufficient to stick to linear models, the latter being conceptually simpler and better understood. Locally stationary linear models (Dahlhaus, 1997; Nason et al., 2000) seem to be a particularly interesting option here, as they combine linearity with a modelling approach whereby the time-varying parameters are modelled as 'well-behaved' functions defined on a compact interval, which enables a meaningful asymptotic estimation theory. Indeed, some authors have applied the locally stationary linear framework to the modelling of log-returns; see for example Fryzlewicz (2005) or Clémençon & Slim (2004), who apply the locally stationary covariance estimation methodology of Donoho et al. (2003) to log-returns.

Motivated by the above discussion, we also follow the 'locally stationary linear' avenue and propose a simple nonstationary model for log-returns in which the time-varying volatility, i.e. the log-return variance, is a piecewise-constant function of time: this enables the modelling of abrupt changes in the stochastic regime which are often expected to follow the arrival of good or bad news in the market. Also, we assume that log-returns at different time points are independent.

2. The nonstationary model and motivation

2·1. *The nonstationary model*

Given a financial instrument $\{P_{t,N}\}_{t=1}^{N}$, such as a stock index, a currency exchange rate or a share price, our object of interest is the log-return series $X_{t,N} := \log(P_{t,N}) - \log(P_{t-1,N})$. We propose the following nonstationary 'stochastic triangular array' model for $\{X_{t,N}\}_{t=1}^{N}$:

$$X_{t,N} = \sigma(t/N) Z_t \quad (t = 1, 2, \ldots, N), \tag{1}$$

where $\sigma(z) : [0, 1] \mapsto \mathbb{R}_+$ is a nonparametric function, and $\{Z_t\}$ is a sequence of independent and identically distributed random variables such that $E(Z_t) = 0$ and $E(Z_t^2) = 1$.

Here, $\sigma(z)$, or alternatively $\sigma^2(z)$, can be viewed as a time-dependent parameter of the proposed model (1), which needs to be estimated from a single stretch of observations $\{X_{t,N}\}$. Note that $\sigma(z)$ is defined over the interval $[0, 1]$, which is common practice in nonparametric regression and is done in order to enable the specification of regularity assumptions for $\sigma(z)$. Indeed, without such regularity assumptions, any attempts at estimating $\sigma(z)$ in a consistent manner would not be possible.

As we are primarily interested in $\sigma^2(z)$, the local variance or volatility of the process $\{X_{t,N}\}_{t=1}^N$, rather than $\sigma(z)$ itself, we now specify the smoothness assumption for $\sigma^2(z)$ which will be used throughout the paper.

Assumption 1. The function $\sigma^2(z)$ is piecewise-constant, bounded from above and away from zero, with a finite but unknown number of jumps.

Additional assumptions on the innovation process $\{Z_t\}$ will be specified later.

In the paper, we demonstrate theoretically and empirically that piecewise stationarity, which is arguably the simplest type of departure from stationarity, is already flexible and powerful enough to enable the successful modelling and forecasting of volatilities. We note that the piecewise-constant modelling of volatilities has also been considered by Mercurio & Spokoiny (2004) and in an unpublished Weierstrass Institute technical report by J. Polzehl and V. Spokoiny.

Our nonparametric approach allows us to avoid the restrictions imposed by the parametric structures of ARCH/GARCH models. Similarly to Stărică & Granger (2005), by modelling volatility as a nonparametric function, we do not claim that random effects do not play any role in the volatility dynamics. In our modelling approach, we express our belief that both past and future returns are manifestations of an unspecified exogeneous economic factor about which we only assume a piecewise-constant nature. Since no obvious candidate for explanatory exogeneous variables is at hand, we model the volatility as a nonparametric function.

We also observe that our approach is different from the use of a piecewise-constant noise variance in the threshold autoregressive models of Tong (1990, Ch. 3). In the latter approach, different autoregressive regimes are followed above and below a certain threshold, which introduces nonlinear dependence in the process. In contrast to that approach, our model is linear. Examples of approaches to fitting piecewise-stationary autoregressive models include Ozaki & Tong (1975) and Kitagawa & Akaike (1978).

2·2. *Explanation of the common properties*

Below, we demonstrate that, provided Assumption 1 holds, the nonstationary model (1) is capable of explaining the most commonly observed properties of log-returns, mentioned in § 1. We introduce the following notation:

$$\bar{X}_N^p = \frac{1}{N}\sum_{t=1}^{N} X_{t,N}^p, \quad \gamma_{X^p}^N(h) = \frac{1}{N}\sum_{t=1}^{N-h} X_{t,N}^p X_{t+h,N}^p - (\bar{X}_N^p)^2.$$

The following proposition holds.

PROPOSITION 1. *Suppose that* $\{X_{t,N}\}_{t=1}^N$ *follows model* (1), *and that Assumption* 1 *holds. Assume further that* $E(Z_t^8) < \infty$. *Suppose that* $\{h_N\}$ *is a sequence such that* $h_N > 0$ *and, for some* $\beta \geqslant 0$, *that* $h_N/N \to \beta$ *as* $N \to \infty$. *Then we have, as* $N \to \infty$,

$$\bar{X}_N^1 \to 0, \tag{2}$$

in mean square,

$$\frac{\bar{X}_N^4}{(\bar{X}_N^2)^2} \to E(Z_t^4)\frac{\int_0^1 \sigma^4(z)dz}{\{\int_0^1 \sigma^2(z)dz\}^2}, \tag{3}$$

in probability,

$$\gamma_{X^1}^N(h) \to 0, \tag{4}$$

in mean square, for a fixed $h > 0$, *and*

$$\gamma^N_{X^2}(h_N) \to \int_0^{1-\beta} \sigma^2(z)\sigma^2(z+\beta)dz - \left\{ \int_0^1 \sigma^2(z)dz \right\}^2, \tag{5}$$

in mean square.

The proofs of results in this paper appear in a technical report by the authors, available at `http://www.maths.bris.ac.uk/~mapzf/hf_vol/hf_vol.pdf`.

Typically, upon assuming that our observations come from a stationary, but not necessarily linear, process, which is indeed often done in log-return analysis, we would use the quantities on the left-hand sides of formulae (2)–(5) as measures of the mean, the kurtosis, the autocovariance and the autocovariance of the squares of the data, respectively. Thus, Proposition 1 demonstrates that the common properties of log-return data, mentioned in § 1, are captured by our model. In particular, note that the ratio on the right-hand side of formula (3) is always greater than 1, unless $\sigma^2(z)$ is constant in which case it is equal to 1. Similarly, if $h_N = h > 0$, then the integral on the right-hand side of formula (5) is always positive, unless $\sigma^2(z)$ is constant in which case it is equal to 0.

The above discussion indicates that care must be taken when applying stationary, global tools to the analysis of log-returns, as the true underlying model might well turn out to be nonstationary, as is indeed the case here. In particular, Proposition 1 demonstrates that the estimated sample autocovariance evaluated under the premise that the process is stationary gives a misleading view as to the true dependence structure of the underlying process. More precisely, it is clear that the true correlations of the process $\{X^2_{t,N}\}^N_{t=1}$ are zero. However, for all $h > 0$, it is straightforward to see from (5) that the autocovariance estimator $\gamma^N_{X^2}(h)$ does not necessarily converge to zero as $N \to 0$ and the correlations appear to persist for all values of h, thus giving the wrong impression that $\{X^2_{t,N}\}^N_{t=1}$ may be correlated, or even have the long-memory property, when in fact they are independent.

We finally note that our model naturally captures the often-observed clustering of volatility. Indeed, the piecewise-constant form of $\sigma^2(z)$ means that the local variance remains at the same level for a number of time units, thus modelling the volatility clustering.

3. A Haar–Fisz estimation theory

3·1. *Motivation*

In this section, we aim to estimate $\sigma^2(t/N)$ at time points $t = 1, 2, \ldots, N$ from a single stretch of observations $\{X_{t,N}\}^N_{t=1}$ from the nonstationary model (1). As we assume $\sigma^2(z)$ to be piecewise-constant, we base our estimator on Haar wavelets, which, being also piecewise-constant, are potentially good 'building blocks' for this purpose. Our estimator uses the principle of nonlinear wavelet shrinkage, thus being potentially well-suited for the estimation of $\sigma^2(z)$ even if it is spatially inhomogeneous; in other words, if the regularity of $\sigma^2(z)$ varies from one region to another. For an overview of wavelet methods in statistics, we refer the reader to the monograph of Vidakovic (1999).

The starting point for these considerations is a reformulation of (1):

$$X^2_{t,N} = \sigma^2(t/N)Z^2_t \quad (t = 1, 2, \ldots, N). \tag{6}$$

Note that $X_{t,N}^2$ is an unbiased but inconsistent estimator of $\sigma^2(t/N)$, and thus needs to be smoothed to achieve consistency. Obviously, (6) can be rewritten as

$$X_{t,N}^2 = \sigma^2(t/N) + \sigma^2(t/N)(Z_t^2 - 1) \quad (t = 1, 2, \ldots, N), \tag{7}$$

so that the problem of estimating $\sigma^2(t/N)$ can be viewed as the problem of removing the 'noise' $\sigma^2(t/N)(Z_t^2 - 1)$ from $\{X_{t,N}^2\}_{t=1}^N$.

Neumann & von Sachs (1995) used a nonlinear wavelet estimation technique in a setting similar to (7). However, their method involved finding an estimator of the local variance of the 'noise', here $\sigma^2(t/N)(Z_t^2 - 1)$, which in our case would amount to finding a pre-estimator of $\sigma^2(t/N)$ itself. This is an obvious drawback of the estimation procedure, and can hamper the practical performance of the method (Fryzlewicz, 2005).

In order to avoid having to find a pre-estimator of $\sigma^2(t/N)$, an obvious step would be to take the logarithmic transformation of (6):

$$\log X_{t,N}^2 = \log \sigma^2(t/N) + \log Z_t^2 \quad (t = 1, 2, \ldots, N). \tag{8}$$

The logarithmic transformation transforms model (6) from multiplicative to additive, and acts as a variance-stabiliser. This setting is similar to the representation of the log-periodogram of a second-order stationary process considered by Wahba (1980). Several authors proposed wavelet techniques for the estimation of the log-periodogram (Moulin, 1994; Gao, 1997), and those techniques could be adapted to our framework. However, any wavelet estimator in the setting specified by (8) would possess two undesirable properties. First, naturally enough, it would be an estimator of $\log \sigma^2(t/N)$, and not of $\sigma^2(t/N)$ itself. Exponentiating this estimator would yield an estimator of $\sigma^2(t/N)$; however, the statistical properties of the latter, such as mean-square consistency, would not be easy to establish; note that, generally, the existence of the second moment of a random variable Y does not imply the existence of the second moment of $\exp(Y)$. Secondly, any wavelet estimator in model (8) would suffer from a bias of order $E(\log Z_t^2)$. Since, as mentioned before, we do not assume any specific distributional form for the innovation process $\{Z_t\}$, the magnitude of the bias correction factor would then be unknown.

In contrast to those unwelcome features, the Haar–Fisz estimation technique which we propose below enjoys the following properties: it uses a variance-stabilising step, which eliminates the need for a local variance pre-estimation, and it yields an asymptotically unbiased, mean-square consistent estimator of $\sigma^2(t/N)$, as opposed to $\log \sigma^2(t/N)$, which removes the need for a bias correction factor.

3·2. *The Haar–Fisz estimation algorithm*

The input to the algorithm is the vector $\{X_{t,N}^2\}_{t=1}^N$: here, we assume that N is an integer power of two; techniques for adapting wavelet transforms to nondyadic sample sizes are described in Wickerhauser (1994). To simplify the notation, we drop the subscript N and consider the sequence $X_t^2 := X_{t,N}^2$. We set $J = \log_2 N$. The estimation algorithm proceeds as follows.

Stage 1. Compute the Haar decomposition of $\{X_t^2\}_{t=1}^N$ using the following algorithm:

(a) let $s_{J,k} := X_k^2$, for $k = 1, 2, \ldots, 2^J$;

(b) for each $j = J-1, J-2, \ldots, 0$, recursively form vectors s_j, d_j and f_j with elements

$$s_{j,k} = \frac{s_{j+1,2k-1} + s_{j+1,2k}}{\surd 2}, \quad d_{j,k} = \frac{s_{j+1,2k-1} - s_{j+1,2k}}{\surd 2}, \quad f_{j,k} = \frac{d_{j,k}}{s_{j,k}},$$

where $k = 1, \ldots, 2^j$.

Stage 2. For each $j = J-1, J-2, \ldots, 0$ and $k = 1, 2, \ldots, 2^j$, let $\mu_{j,k} := E(d_{j,k})$. For most levels j, in a sense to be made precise later, estimate $\mu_{j,k}$ by

$$\hat{\mu}^{(h)}_{j,k} = s_{j,k} f_{j,k} I(|f_{j,k}| > t_j) = d_{j,k} I(|f_{j,k}| > t_j), \qquad (9)$$

corresponding to hard thresholding, or by

$$\hat{\mu}^{(s)}_{j,k} = s_{j,k}\, \mathrm{sgn}(f_{j,k})(|f_{j,k}| - t_j)_+, \qquad (10)$$

corresponding to soft thresholding, where $I(.)$ and sgn(.) are the indicator and sign functions, respectively, and $(x)_+ = \max(0, x)$. In other words, we 'kill' each $d_{j,k}$ if and only if the corresponding 'Haar–Fisz coefficient' $f_{j,k}$ does not exceed in absolute value a certain threshold t_j, to be specified later. Note that this is different from classical wavelet thresholding in that the thresholded quantity $d_{j,k}$ and the 'thresholding statistic' $f_{j,k}$ are different.

Stage 3. Invert the Haar decomposition in the usual way to obtain an estimate of $\sigma^2(t/N)$ at time points $t = 1, 2, \ldots, N$. Call the resulting estimate $\hat{\sigma}^2_{(h)}(t/N)$, for hard thresholding, or $\hat{\sigma}^2_{(s)}(t/N)$, for soft thresholding. Explicit formulae for these two estimators are given later in this section.

Asymptotic Gaussianity and variance stabilisation for certain random variables of the form $(X - Y)/(X + Y)$, where X and Y are nonnegative, independent random variables, were studied by Fisz (1955): hence we label the $f_{j,k}$ the 'Haar–Fisz coefficients'. The main heuristic idea here is that the variance of $f_{j,k}$, for most j, k, does not depend on $\sigma^2(z)$. Consider the following example, with $j = J-1$ and $k = 1$. The Haar–Fisz coefficient $f_{J-1,1}$ has the form

$$f_{J-1,1} = \frac{X_1^2 - X_2^2}{X_1^2 + X_2^2} = \frac{\sigma^2(1/N)Z_1^2 - \sigma^2(2/N)Z_2^2}{\sigma^2(1/N)Z_1^2 + \sigma^2(2/N)Z_2^2}.$$

Suppose now that $\sigma^2(1/N) = \sigma^2(2/N)$; this is likely as $\sigma^2(z)$ is piecewise-constant. We then have $f_{J-1,1} = (Z_1^2 - Z_2^2)/(Z_1^2 + Z_2^2)$, and the variance of $f_{J-1,1}$ does not depend on $\sigma^2(z)$. Thus, the thresholds t_j in (9) and (10) also do not need to depend on $\sigma^2(z)$, and can therefore be selected more easily.

In the above example, if $\sigma^2(1/N)$ were not equal to $\sigma^2(2/N)$, that is if a jump occurred between times $1/N$ and $2/N$, then the distribution of $f_{J-1,1}$ would depend on $\sigma^2(1/N)$ and $\sigma^2(2/N)$ in a nontrivial way. In particular, we could expect $f_{J-1,1}$ to be significantly deviated from zero, if the value of $\sigma^2(1/N)$ was much different from that of $\sigma^2(2/N)$. In that case the corresponding coefficient $d_{J-1,1}$ ought to 'survive' the process of thresholding.

Note that the Haar–Fisz transform for Poisson data, an algorithmic device for stabilising the variance of Poisson data and bringing their distribution closer to normality, was introduced by Fryzlewicz & Nason (2004).

We now give precise and explicit definitions of $\hat{\sigma}^2_{(h)}(t/N)$ and $\hat{\sigma}^2_{(s)}(t/N)$ in terms of Haar wavelet vectors. For $j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$, define the Haar wavelet vectors $\{\psi_{j,k}(t)\}_{t=1}^{2^J}$ as

$$\psi_{j,k}(t) = 2^{(j-J)/2} I\left[t \in \left\{(k-1)2^{J-j} + 1, \ldots, \left(k - \frac{1}{2}\right)2^{J-j}\right\}\right]$$
$$- 2^{(j-J)/2} I\left[t \in \left\{\left(k - \frac{1}{2}\right)2^{J-j} + 1, \ldots, k2^{J-j}\right\}\right].$$

Fix $\delta \in (0, 1)$. For each $N = 2^J$, define the set $\mathscr{I}_N = \{(j, k) : j < J^*\}$, with $2^{J^*} = O(N^{1-\delta})$. The estimators $\hat{\sigma}^2_{(h)}(t/N)$ and $\hat{\sigma}^2_{(s)}(t/N)$ are defined as

$$\hat{\sigma}^2_{(h)}(t/N) = \bar{X}^2_N + \sum_{(j,k) \in \mathscr{I}_N} \hat{\mu}^{(h)}_{j,k} \psi_{j,k}(t), \tag{11}$$

$$\hat{\sigma}^2_{(s)}(t/N) = \bar{X}^2_N + \sum_{(j,k) \in \mathscr{I}_N} \hat{\mu}^{(s)}_{j,k} \psi_{j,k}(t), \tag{12}$$

where $\hat{\mu}^{(h)}_{j,k}$ and $\hat{\mu}^{(s)}_{j,k}$ are as in formulae (9) and (10), respectively, with

$$t_j = 2^{-(J-j-1)/2} \sqrt{(2 \log N)}. \tag{13}$$

Outside the set $\mathscr{I}_N$, we simply define $\hat{\mu}^{(h)}_{j,k} = \hat{\mu}^{(s)}_{j,k} = 0$. Let

$$v := E(|Z^2_t - 1|^2), \tag{14}$$

and consider the following assumption.

Assumption 2. The distribution of the random variable Z^2_t has no atom at 0, and there exist $c > 0$ and $\gamma \geqslant 0$ such that $E(|Z^2_t - 1|^n) \leqslant c^{n-2} (n!)^{1+\gamma} v$, for all $n \geqslant 3$.

Remark 1. By elementary properties of the Gaussian and Laplace distributions (Johnson et al., 1994, Ch. 13, 24), Assumption 2 is satisfied, in particular, if Z_t is standard Gaussian, with $v = 2$ and $\gamma = 0$, or standard Laplace, with $v = 5$ and $\gamma = 2$. Assumption 2 can also accommodate other distributions which are leptokurtic or possess a degree of skewness.

The following theorem, proved in the Appendix, demonstrates the mean-square consistency of $\hat{\sigma}^2_{(h)}(z)$ and $\hat{\sigma}^2_{(s)}(z)$.

THEOREM 1. *Suppose that* $\{X_{t,N}\}_{t=1}^N$ *follows model* (1), *and that Assumptions* 1–2 *hold. Let* (e) *be either one of* (h) *and* (s). *Then we have*

$$\begin{aligned}\frac{1}{N} \sum_{t=1}^{N} E\left\{\hat{\sigma}^2_{(e)}\left(\frac{t}{N}\right) - \sigma^2\left(\frac{t}{N}\right)\right\}^2 &= \frac{v}{N^2} \sum_{t=1}^{N} \sigma^4\left(\frac{t}{N}\right) + \frac{1}{N} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} E(\hat{\mu}^{(e)}_{j,k} - \mu_{j,k})^2 \\ &= O(N^{-\min(1-\delta,\, 2/v)}),\end{aligned} \tag{15}$$

where v *is defined as in formula* (14).

Remark 2. Note that, if Z_t is standard Gaussian, so that $v = 2$, the mean-square error rate in (15) reduces to $O(N^{-1+\delta})$, which is arbitrarily close to the parametric rate of $O(N^{-1})$. Intuitively, this is not surprising as our problem is ‘almost parametric’ in the sense that our target function is piecewise-constant with a finite number of jumps, but the exact number, locations or magnitudes of the jumps are not known. It is clear from the proof of Theorem 1 that the exact parametric rate is unattainable for our estimation procedure, because we only use nontrivial estimators of $\mu_{j,k}$ in the set $\mathscr{I}_N$, which is essential for a certain asymptotic normality effect to hold. This effect holds for any choice of $\delta > 0$, and thus, in theory, it is beneficial to choose δ to be 'as small as possible'. If we employ the asymptotic normality as a tool, the rate is shown to be $O(N^{-1+\delta})$ given that the target function is piecewise-constant.

3·3. *Noise-free reconstruction*

In this section, we consider the case when the errors Z_t in (1) are standard Gaussian. We construct an estimator of $\sigma^2(z)$ which possesses the following noise-free reconstruction property: if the true function $\sigma^2(z)$ is a constant function of z, then, with high probability, our estimator of $\sigma^2(z)$ is also constant and equal to the empirical mean of $\{X_{t,N}^2\}_{t=1}^N$.

The noise-free reconstruction property guarantees that estimators obtained using our method have a visually appealing character and do not exhibit spurious 'spikes', even for a nonconstant $\sigma^2(z)$. This is achieved by requiring that, asymptotically, no pure 'noise' coefficient survives the thresholding procedure.

For the noise-free reconstruction property to hold, we require that the probability of any $f_{j,k}$ exceeding $\tilde{t}_j$ should tend to 0 as $N \to \infty$, or, to be more precise,

$$\text{pr}\left\{\bigcup_{j=0}^{J-1}\bigcup_{k=1}^{2^j}(|f_{j,k}| > \tilde{t}_j)\right\} \to 0, \tag{16}$$

as $N \to \infty$, where $J = \log_2 N$, $f_{j,k}$ is the Haar–Fisz coefficient of $\{X_{t,N}\}$ and $\tilde{t}_j$ are appropriately chosen thresholds. We note that the thresholds used in this case are different from those given in § 3·2.

In order to derive the appropriate thresholds $\tilde{t}_j$ we use the following lemma.

Lemma 1. *Let $\{X_i\}_{i=1}^{2m}$ be a sequence of independent and identically distributed χ_1^2 random variables, and let $X^{(1)} = \sum_{i=1}^m X_i$ and $X^{(2)} = \sum_{i=m+1}^{2m} X_i$. Then the ratio*

$$(X^{(1)} - X^{(2)})/(X^{(1)} + X^{(2)})$$

is distributed as $2Y - 1$, where $Y \sim \text{Be}(m/2, m/2)$.

We now derive the thresholds $\tilde{t}_j$. As the distribution of $f_{j,k}$ does not depend on k, we can define $\alpha_j(N) = \text{pr}(|f_{j,k}| < \tilde{t}_j)$. We have the following bound for (16), based on the Bonferroni inequality:

$$\text{pr}\left\{\bigcup_{j=0}^{J-1}\bigcup_{k=1}^{2^j}(|f_{j,k}| > \tilde{t}_j)\right\} \leqslant \sum_{j=0}^{J-1} 2^j\{1 - \alpha_j(N)\}. \tag{17}$$

Our objective is to choose $\{\alpha_j(N)\}_{j=0}^{J-1}$ such that $\sum_{j=0}^{J-1} 2^j\{1-\alpha_j(N)\} \to 0$, as $N \to \infty$. The choice of $\{\alpha_j(N)\}_{j=0}^{J-1}$ will determine how fast $\sum_{j=0}^{J-1} 2^j\{1-\alpha_j(N)\}$ approaches zero and thus the rate of convergence. Probably the simplest option is to mimick standard universal thresholding in a classical independent and identically distributed Gaussian nonparametric regression setting for wavelets, where the analogue of $\alpha_j(N)$ is constant across scales, that is $\alpha_j(N) = \alpha(N)$, and the rate of convergence to 0 of the probability corresponding to (16) is equal to $(\pi J \log 2)^{-\frac{1}{2}}$. To guarantee such a rate we choose $\alpha(N)$ such that

$$\text{pr}\left\{\bigcup_{j=0}^{J-1}\bigcup_{k=1}^{2^j}(|f_{j,k}| > \tilde{t}_j)\right\} \leqslant \sum_{j=0}^{J-1} 2^j\{1 - \alpha(N)\} = \frac{1}{\surd(\pi J \log 2)},$$

which is solved uniquely by

$$\alpha^*(N) = 1 - (2^J - 1)^{-1}(\pi J \log 2)^{-\frac{1}{2}}. \tag{18}$$

In the light of what was said above, $\alpha^*(N)$ guarantees the convergence of (16) to 0 at a rate of at least $\{\pi J \log(2)\}^{-\frac{1}{2}}$.

By Lemma 1, $f_{j,k}$ has a $2\mathrm{Be}(2^{J-j-2}, 2^{J-j-2})-1$ distribution, and $\tilde{t}_j$'s are now easily found numerically by solving $\alpha^*(N)=\mathrm{pr}(|f_{j,k}|<\tilde{t}_j)$. We note that noise-free reconstruction is also possible for random variables Z_t which have distributions other than Gaussian, so long as the exact distribution of $(X^{(1)}-X^{(2)})/(X^{(1)}+X^{(2)})$ is known, where $X^{(1)}=\sum_{i=1}^{m} Z_i^2$ and $X^{(2)}=\sum_{i=m+1}^{2m} Z_i^2$.

Figure 1 compares the thresholds t_j and $\tilde{t}_j$ for $J=10$ and $j=0,\ldots,J-1$. Note that t_j's exceed 1 at the 4 finest scales, and therefore no coefficient at these scales survives the thresholding; remember that $|f_{j,k}|$ is always bounded from above by 1.

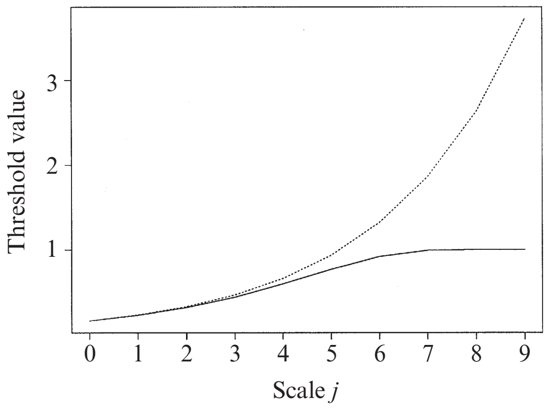

Fig. 1. Thresholds $\tilde{t}_j$, solid line, and t_j, dotted line, for $J=10$ and $j=0,\ldots,J-1$. See discussion in § 3·3.

In classical Gaussian wavelet regression, some authors argue that, instead of modelling the analogue of $\alpha_j(N)$ as constant across scales, one can obtain more accurate estimators by allowing it to decrease from finer to coarser scales (Antoniadis & Fryzlewicz, 2006). For simplicity, we consider a linear dependence of $\alpha_j(N)$ on j:

$$\alpha_j(N)=\alpha_{J-1}(N)\frac{j}{J-1}+\alpha_0(N)\frac{J-1-j}{J-1}.$$

The equation for $(\alpha_0(N), \alpha_{J-1}(N))$ is

$$c_J=\sum_{j=0}^{J-1} 2^j\left\{1-\alpha_{J-1}(N)\frac{j}{J-1}-\alpha_0(N)\frac{J-1-j}{J-1}\right\},$$

where $c_J \downarrow 0$ is the desired rate of convergence. This simplifies to

$$\alpha_{J-1}(N)\{2^J(J-2)+2\}+\alpha_0(N)(2^J-J-1)=(2^J-1-c_J)(J-1).$$

One possibility is to set $\alpha_{J-1}(N)=\alpha^*(N)$ and then solve for $\alpha_0(N)$. As a special case, note that setting $\alpha_{J-1}(N)=\alpha^*(N)$ and $c_J=(\pi J \log 2)^{-\frac{1}{2}}$ gives the solution $\alpha_0(N)=\alpha^*(N)$, which implies that $\alpha_j(N)=\alpha^*(N)$ for all j; in this case, $\alpha_j(N)$ does not depend on j.

4. Currency exchange rate examples

In this section, we exhibit the performance of various versions of our Haar–Fisz volatility estimator on two currency-exchange datasets, namely the logged and differenced daily exchange rates between the U.S. dollar and the British pound and between the dollar and the Japanese yen, both running from 1 January 1990 to 31 December 1999. The data are available from the U.S. Federal Reserve website `http://www.federalreserve.gov/releases/h10/Hist/default1999.htm`. We have also tested our estimator on other exchange rate datasets available from the above website but for lack of space we only provide graphical illustration of its performance on the dollar/pound and dollar/yen series in this section. However, the discussion below applies to all of the exchange rate time series available from the above website.

The length of each series is $n = 2515$, but, as our estimators require the length of input to be a power of two, we only consider the last $N = 2048$ observations in both series, so that $J = \log_2 N = 11$. Those are plotted in Fig. 2.

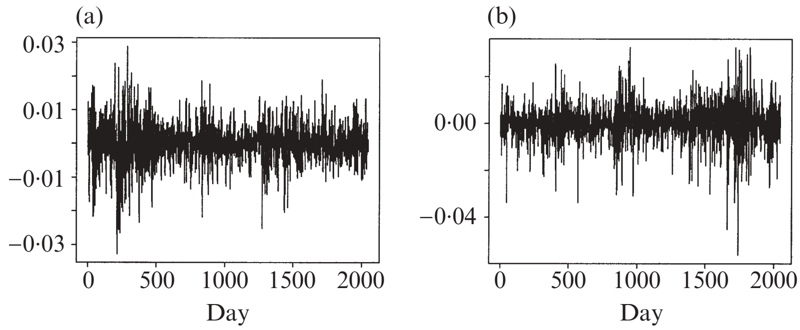

Fig. 2. The last $N = 2048$ observations of (a) the dollar/pound and (b) the dollar/yen logged and differenced exchange rate series, running from 8 November 1991 to 31 December 1999.

We now single out a few specific versions of our Haar–Fisz volatility estimator.

Method MS-H: Our Haar–Fisz algorithm with hard thresholding and thresholds t_j, see formula (13), which guarantee mean-square consistency. We take $J^* = J - 1$; see § 3·2 for details.

Method NF-p-H(-TI): Our Haar–Fisz algorithm with hard thresholding and noise-free reconstruction thresholds $\tilde{t}_j$ chosen in such a way that $\alpha_{J-1} = \alpha^*$, see formula (18), and $\alpha_0 = (p/100)\alpha_{J-1}$, where $p \leqslant 100$; see § 3·3 for details. The acronym TI denotes the translation-invariant version: in translation-invariant versions of wavelet-based denoising algorithms, the final estimator is obtained as the average of the estimators obtained for all circular shifts of the data. This is common practice in wavelet regression. The fast $O(N \log N)$ implementation of translation-invariant wavelet thresholding algorithms uses the non-decimated wavelet transform (Nason & Silverman, 1995).

Method NF-p-S(-TI): The same as above, with soft thresholding.

We have tested several versions of our estimator by looking at the behaviour of empirical residuals for each of the datasets. To be more specific, let $X_{t,N} = \sigma(t/N)Z_t$ denote a series of currency-exchange log-returns, and let $\hat{\sigma}^2(t/N)$ be any Haar–Fisz estimator of $\sigma^2(t/N)$. We define the empirical residuals as $\hat{Z}_t = X_{t,N}/\hat{\sigma}(t/N)$. We are satisfied with the performance of $\hat{\sigma}^2(t/N)$ if the sequence $\hat{Z}_t$ looks 'stationary' and displays only very little autocorrelation in the squares; that is the p-value of the Ljung–Box test for lack of serial correlation in $\hat{Z}_t^2$ is above a prespecified threshold λ, with $\lambda = 0{\cdot}05$ in all of the examples in the paper.

In an extensive empirical study which compared several parameter configurations for several currency exchange datasets, we found that, for $p = 100$, the corresponding estimators NF-100-H and NF-100-S(-TI) often oversmoothed, in the sense that the empirical residuals displayed significant dependence in the squares. The suitable value of p was then chosen by decreasing p over the grid 100, 99, 98, ... until the Ljung–Box test indicated no significant correlation in the squared empirical residuals. We found that either of the two estimators NF-98-S or NF-97-S, as well as their TI versions, performed well for most of the datasets considered. On the other hand, the estimators NF-p-H for $p < 100$ were often extremely 'spiky'. Typically, NF-100-H-TI produced correctly behaved empirical residuals, although the reconstructions were also often spiky. The price for using any translation-invariant estimator was that, naturally enough, we lost the piecewise-constant nature of the reconstructions and increased the computational effort from $O(N)$ to $O(N \log N)$.

Furthermore, we found that the MS-H estimator typically gave slightly oversmoothed reconstructions. This was because, as mentioned in § 3·3, the thresholds t_j were larger than 1 at the four finest scales, which meant that no detail coefficient $d_{j,k}$ at those scales survived the thresholding.

To summarise, NF-p-S, NF-p-S-TI and NF-p-H-TI were the preferred estimators. For the dollar/pound series, the corresponding values of p for those estimators, selected by the automatic procedure described above, were $p = 97$, $p = 97$ and $p = 100$, respectively. The respective p-values of the Ljung–Box test were 0·09, 0·06 and 0·82. For the dollar/yen series, the selected values of p for the above three estimators were $p = 97$, $p = 98$ and $p = 100$, respectively. The respective p-values of the Ljung–Box test were 0·19, 0·18 and 0·94.

Figures 3(a), (c) and (e) show the results for the dollar/pound series. Figure 3(a) shows the NF-97-S-TI estimate and Fig. 3(c) shows the NF-100-H-TI estimate. The NF-97-S estimate is a piecewise-constant function whose breakpoints can be loosely interpreted as 'significant changes' in the volatility. Figure 3(e) shows the squared returns and the locations of the break points of NF-97-S. For clarity, we only plot the first 250 observations, which roughly corresponds to one business year starting on 8 November 1991. Figures 3(b), (d) and (f) show the corresponding results for the dollar/yen series except that Fig. 3(b) shows the NF-98-S-TI estimate and Fig. 3(f) shows the last 250 observations, which roughly corresponds to one business year starting on 1 January 1999. Note that translation-invariant estimates are not useful for break-point detection as they are continuous.

Although the soft-thresholding translation-invariant estimates were smoother and thus more visually appealing than the hard-thresholding translation-invariant estimates, it was far from obvious that they should be preferred, as the p-values for the latter were much higher. This might imply that some of the spikes observed in the hard-thresholding estimates were not merely artefacts from hard thresholding but served to explain significant transient features of the volatility function. On the other hand, the low p-values produced by the soft-thresholding estimates might be due to the occasional bias introduced by soft thresholding, which is a well-known phenomenon in the classical Gaussian regression context.

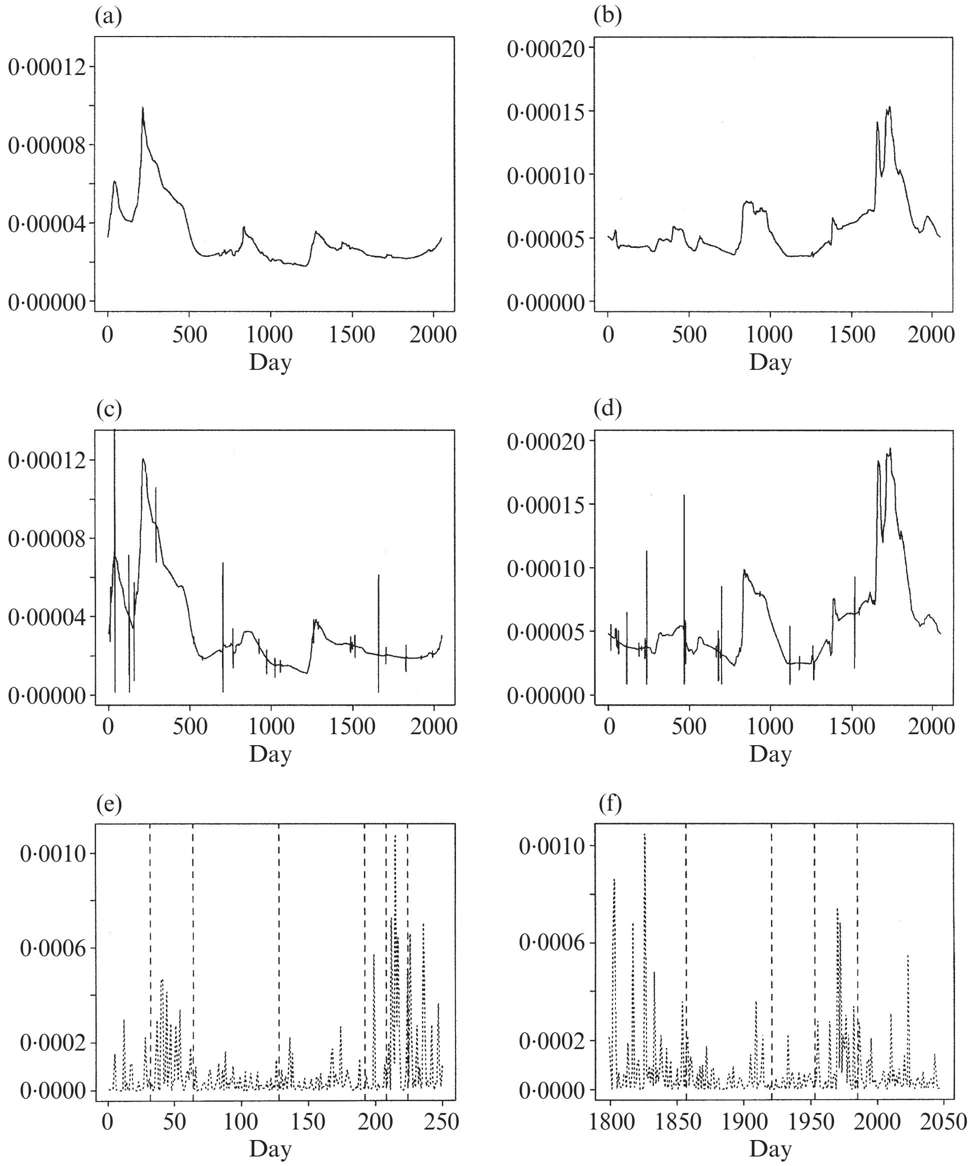

Fig. 3. (a), (c) and (e) show empirical results for the dollar/pound series; (b), (d) and (f) show the dollar/yen series. (a) NF-97-S-TI estimate; (b) NF-98-S-TI estimate; (c) NF-100-H-TI estimate; (d) NF-100-H-TI estimate; (e) first 250 observations of the squared series, dotted, and the corresponding breakpoints of the NF-97-S estimate, dashed; (f) last 250 observations of the squared series, dotted, and the corresponding breakpoints of the NF-97-S estimate, dashed.

5. Forecasting currency exchange rate volatility

In this section, we describe the outcome of an empirical study designed to assess the forecasting ability of our model and compare it to that of the benchmark stationary GARCH(1, 1) process with Gaussian innovations, as well as a simple 'moving window' procedure. Suppose that we observe $X^2_{1,N}, X^2_{2,N}, \ldots, X^2_{t,N}$ from model (1) and want to forecast the volatility at times $t+1, \ldots, t+h$, where $t+h \leqslant N$. The mean-square-optimal forecasts are given by

$$\sigma^{2,\mathrm{HF}}_{t|t+h,N} := E(X^2_{t+h,N} | X^2_{1,N}, \ldots, X^2_{t,N}) = \sigma^2\left(\frac{t+h}{N}\right).$$

Obviously, the true value of $\sigma^2\{(t+h)/N\}$ is unknown at time t and the best that we can do is to extrapolate it as $\sigma^{2,\mathrm{HF}}_{t|t+h,N} = \hat{\sigma}^2(t/N)$, where $\hat{\sigma}^2(t/N)$ is any of our Haar–Fisz estimates of $\sigma^2(t/N)$. Note that contrary to the GARCH case (Bera & Higgins, 1993) our forecasts do not depend on the forecasting horizon. In the examples below, we use the versions NF-98-S and NF-100-S of our estimator to compute $\hat{\sigma}^2(t/N)$ for the above forecasts. Both estimates are computed from $X^2_{t-1023,N}, \ldots, X^2_{t,N}$. Then, for all h, the forecast $\sigma^{2,\mathrm{HF}}_{t|t+h,N}$ is the value of our estimate at the last time point t/N.

For GARCH-based forecasts, we forecast the volatility at time $t+1, \ldots, t+h$ from $X^2_1, \ldots, X^2_t$ using the following methods.

Method G-SCROLL: We fit the stationary GARCH$(1, 1)$ model with standard Gaussian innovations to $X_{t-1023}, \ldots, X_t$ using the S-Plus routine `garch`, and then forecast the volatility using the S-Plus routine `predict`.

Method G-NSCROLL: This is similar to G-SCROLL, but the model is fitted to $X_1, \ldots, X_t$.

For the simple moving window procedure, labelled MW, the predicted volatility at times $t+1, \ldots, t+h$ is the empirical mean of the values $(X^2_{t-h+1}, \ldots, X^2_t)$. Here, we use the rule of thumb advocated by Hull (1997, p. 233) whereby the time period over which volatility is estimated should be set equal to the time period h over which it is to be applied.

Let $\sigma^2_{t|t+h}$ denote a generic volatility forecast, computed using any of the above procedures. For each of the currency-exchange datasets tested, for which details are given below, we compute the following error measure: for each $t = 1024, \ldots, N-250$, where N, the length of each dataset, oscillates around 2500 but varies from one dataset to another, we compute the quantity $\bar{\sigma}^2_{t|t+250} = \sum_{h=1}^{250} \sigma^2_{t|t+h}$, and compare it to the 'realised' volatility $\bar{X}^2_{t|t+250} = \sum_{h=1}^{250} X^2_{t+h}$, using the average squared error

$$\mathrm{ASE}_{250,1024,N} = \frac{1}{N-1273} \sum_{t=1024}^{N-250} (\bar{\sigma}^2_{t|t+250} - \bar{X}^2_{t|t+250})^2.$$

In other words, we forecast the volatility one business year of 250 days ahead; this is done to compare the long-term forecasting abilities of the competitors.

Our datasets are logged and differenced currency exchange rates between the U.S. dollar and a variety of other currencies, available from the web address given in § 4. The other currencies in Table 1 are Australia dollar, Canada dollar, Switzerland franc, Denmark kroner, United Kingdom pound, Hong Kong dollar, Japan yen, South Korea Won, Norway kroner, New Zealand dollar, Sweden kronor, Singapore dollar, Thailand baht, Taiwan new dollar and South Africa rand. Table 1 lists the values, scaled by the number in the right-most column and rounded, of ASE attained by the competing methods for each currency.

The gaps in Table 1 indicate cases in which the stationary GARCH$(1, 1)$ model failed to fit at several points of the corresponding exchange-rate series, producing forecasts which were extremely inaccurate. This was because the numerical maximiser of the likelihood in the S-Plus routine `garch` failed to converge. Our NF-100-S method and the simple MW technique performed the best, or nearly the best, for 7 out of the 15 datasets, and are

Table 1: *Currency exchange rate example. Values of* ASE *for long-term forecasts using the methods described in* § 5. *The best results, and those within* 10% *of the best ones, are in italics*

Currency	G-NSCROLL	G-SCROLL	MW	NF-98-S	NS-100-S	Scaling
AUD	3122	3673	*2944*	*2776*	3095	10^8
CAD	*1610*	2280	2047	1955	*1770*	10^9
CHF	*2844*	*2819*	5168	4574	3541	10^8
DKK	2183	*1658*	2530	*1697*	*1746*	10^8
GBP	23139	15192	7796	*6957*	8645	10^9
HKD	—	—	*3038*	4740	3841	10^{13}
JPY	*1060*	*1006*	1213	1130	*1056*	10^8
KRW	2826	—	*1333*	1938	1657	10^5
NOK	3634	2812	2173	2885	*1846*	10^8
NZD	14502	10414	*9293*	*8847*	10849	10^8
SEK	4451	3812	*741*	1309	1706	10^8
SGD	*4709*	142174	6414	5686	*5122*	10^8
THB	—	—	*3806*	4127	*3580*	10^6
TWD	—	—	4419	4773	*3953*	10^8
ZAR	—	—	*2017*	3321	2600	10^7

AUD, Australian dollar; CAD, Canada dollar; CHF, Switzerland franc; DKK, Denmark kroner; GBP, United Kingdom pound; HKD, Hong Kong dollar; JPY, Japan yen; KRW, South Korea won; NOK, Norway kroner; NZD, New Zealand dollar; SEK, Sweden kronor; SGD, Singapore dollar; THB, Thailand baht; TWD, Taiwan new dollar; ZAR, South Africa rand.

clearly the two preferred options here. Either of our two methods performed the best, or nearly the best, for 10 out of the 15 datasets.

In practice, our recommendation is to use our forecasting technique based on our Haar–Fisz estimation method with soft thresholding and noise-free reconstruction thresholds, where $\alpha_{J-1} = \alpha^*$ and α_0 is chosen from a pre-set grid $\{\alpha_{0,l}\}_{l=1}^{L}$ by comparing the performance of the method on the observed part of the series and choosing the value of $\alpha_{0,l}$ which performs the best. We have found that

$$\{\alpha_{0,l}\}_{l=1}^{5} = \left\{\frac{95+l}{100}\alpha_{J-1}\right\}_{l=1}^{5}$$

is a good practical choice for the grid.

In our empirical study, we have also found that forecasts based on our Haar–Fisz estimation technique with hard thresholding tend to be less accurate and therefore their use is not recommended.

Finally we consider short-term volatility forecasting. It is well known that, even in the simplest case of the stationary Gaussian GARCH(1, 1) model, the GARCH framework provides excellent short-term volatility forecasts. In a brief simulation study, we have compared the performance of our NF-98-S and NF-100-S algorithms to that of the G-NSCROLL technique in forecasting one-day-ahead volatility of the above datasets, except for the five series for which GARCH(1, 1) does not give a good fit, as explained above. We have found that the ratio of ASE for the worst of our two algorithms to ASE for G-NSCROLL ranged between 0·99 and 1·09 for all of the datasets, which demonstrates good performance of our technique also in the case of short-term forecasts.

Acknowledgement

Piotr Fryzlewicz thanks Theofanis Sapatinas and David Hand for financial support and Theofanis Sapatinas for his hospitality during a visit to Nicosia. He is also grateful to the Department of Mathematics, Imperial College, London, where most of the work for the current version of this paper was done, for the excellent research environment. Theofanis Sapatinas was partly supported by a University of Cyprus 2004 Research Grant. Suhasini Subba Rao was supported by a grant from the Deutsche Forschungsgemeinschaft. We thank the editor, an associate editor, a referee, Guy Nason and Rainer von Sachs for their helpful comments.

Appendix

Properties of the Haar–Fisz estimator

Proof of Theorem 1. First we consider the case of $(e)=(h)$. The first of the two equalities in (15) arises from the orthonormality of the discrete Haar transform. Note that the term $vN^{-2}\sum_{t=1}^{N}\sigma^4(t/N)$ arises because of the inclusion of the term $\bar{X}_N^2$ in the estimator (11). We now show the second equality. For notational clarity, let $d_{1,j,k}=s_{j+1,2k-1}/\surd 2$ and $d_{2,j,k}=s_{j+1,2k}/\surd 2$, so that $d_{j,k}=d_{1,j,k}-d_{2,j,k}$ and $s_{j,k}=d_{1,j,k}+d_{2,j,k}$. Also let $\mu_{i,j,k}=E(d_{i,j,k})$ and $w_{i,j,k}^2=\text{var}(d_{i,j,k})$ for $i=1,2$. Finally, let $w_{j,k}^2=\text{var}(d_{j,k})$.

For the reader's convenience, we now give explicit formulae for $d_{1,j,k}$ and $d_{2,j,k}$:

$$d_{1,j,k}=2^{(j-J)/2}\sum_{i=2^{J-j}(k-1)+1}^{2^{J-j}(k-1/2)}X_i^2,\quad d_{2,j,k}=2^{(j-J)/2}\sum_{i=2^{J-j}(k-1/2)+1}^{2^{J-j}k}X_i^2.$$

We now compute the risk of $\hat{\mu}_{j,k}^{(h)}$ for $(j,k)\in\mathscr{J}_N$.

Case 1. Here $\sigma^2(i/N):=\text{const}:=\sigma^2$, for $i=2^{J-j}(k-1)+1,\ldots,2^{J-j}k$, so that $\mu_{1,j,k}=\mu_{2,j,k}$. Without loss of generality, consider $k=1$ to shorten the notation.

$$\begin{aligned}E&\left\{(d_{1,j,1}-d_{2,j,1})I\left(\frac{|d_{1,j,1}-d_{2,j,1}|}{d_{1,j,1}+d_{2,j,1}}>t_j\right)-(\mu_{1,j,1}-\mu_{2,j,1})\right\}^2\\&=2^{j-J}\sigma^4E\left[\left\{\sum_{i=1}^{2^{J-j-1}}(Z_i^2-Z_{i+2^{J-j-1}}^2)^2+\sum_{\substack{i,l=1\\ i\neq l}}^{2^{J-j-1}}(Z_i^2-Z_{i+2^{J-j-1}}^2)(Z_l^2-Z_{l+2^{J-j-1}}^2)\right\}\right.\\&\qquad\left.\times I\left(\frac{|d_{1,j,1}-d_{2,j,1}|}{d_{1,j,1}+d_{2,j,1}}>t_j\right)\right].\end{aligned}\tag{A1}$$

Note that, by symmetry arguments, for any $i\neq l$, we have

$$E\left\{(Z_i^2-Z_{i+2^{J-j-1}}^2)(Z_l^2-Z_{l+2^{J-j-1}}^2)I\left(\frac{|d_{1,j,1}-d_{2,j,1}|}{d_{1,j,1}+d_{2,j,1}}>t_j\right)\right\}=0,$$

which simplifies (A1) to

$$\begin{aligned}&\frac{\sigma^4}{2}E\left\{(Z_1^2-Z_{1+2^{J-j-1}}^2)^2I\left(\frac{|d_{1,j,1}-d_{2,j,1}|}{d_{1,j,1}+d_{2,j,1}}>t_j\right)\right\}\\&\qquad\leqslant\frac{\sigma^4}{2}[E\{(Z_1^2-Z_{1+2^{J-j-1}}^2)^{2r}\}]^{1/r}\,\text{pr}\left(\frac{|d_{1,j,1}-d_{2,j,1}|}{d_{1,j,1}+d_{2,j,1}}>t_j\right)^{1-1/r},\end{aligned}\tag{A2}$$

where the above step uses Hölder's inequality with $r > 1$ but otherwise arbitrary. Simple algebra gives

$$\text{pr}\left(\frac{|d_{1,j,1} - d_{2,j,1}|}{d_{1,j,1} + d_{2,j,1}} > t_j\right) = 2\,\text{pr}\left(\frac{d_{1,j,1} - d_{2,j,1}}{d_{1,j,1} + d_{2,j,1}} > t_j\right)$$

$$= 2\,\text{pr}\left[\frac{2^{(j-J)/2}}{\surd\{v(1+t_j^2)\}}\left\{\sum_{i=1}^{2^{J-j-1}} (Z_i^2 - 1)(1-t_j) - \sum_{i=2^{J-j-1}+1}^{2^{J-j}} (Z_i^2-1)(1+t_j)\right\} > \frac{t_j 2^{(j-J)/2}}{\surd\{v(1+t_j^2)\}}\right]. \quad \text{(A3)}$$

Since the condition of Theorem 1 from Rudzkis et al. (1978) holds, because of our Assumption 2, we are able to apply the Corollary to Theorem 1 from Rudzkis et al. (1978). If we recall that $t_j = 2^{(-J+j+1)/2}\surd(2\log N)$, that $t_j \to 0$ on $\mathscr{J}_N$, and that $2^j < 2^{J^*} = O(N^{1-\delta})$, it is easy to see that

$$\frac{t_j 2^{(J-j)/2}}{\surd\{v(1+t_j^2)\}} = o\left\{\left(\frac{2^{(J-j)/2}[\surd\{v(1+t_j^2)\}]/(1+t_j)}{2\max\{c, \surd v\}}\right)^v\right\},$$

as $N \to \infty$, for any positive v. Therefore, by the Corollary to Theorem 1 from Rudzkis et al. (1978), we bound (A3) from above by

$$2^{(-J+j+1)/2} 2C\,\text{pr}\left[N(0,1) > \frac{t_j 2^{(J-j)/2}}{\surd\{v(1+t_j^2)\}}\right]. \quad \text{(A4)}$$

Recalling again that $t_j = 2^{(-J+j+1)/2}\surd(2\log N)$ and denoting by $\Phi(.)$ the cumulative distribution function of a $N(0,1)$ random variable, we now bound (A4) from above using the Mill's ratio inequality (Shorack & Wellner, 1986, p. 850):

$$2C\left(1 - \Phi\left[\frac{2\surd(\log N)}{\surd\{v(1+t_j^2)\}}\right]\right) \leqslant C\exp\left\{-\frac{4\log N}{2v(1+t_j^2)}\right\} = CN^{-2/\{v(1+t_j^2)\}}. \quad \text{(A5)}$$

Letting $C_r = [E\{(Z_1^2 - Z_{1+2^{J-j-1}}^2)^{2r}\}]^{1/r}$ and inserting (A5) in (A2), we obtain

$$\frac{\sigma^4}{2} C_r\,\text{pr}\left(\frac{|d_{1,j,1} - d_{2,j,1}|}{d_{1,j,1} + d_{2,j,1}} > t_j\right)^{1-1/r} \leqslant \sigma^4 \tilde{C}_r N^{-\{2(1-1/r)\}/\{v(1+t_j^2)\}}, \quad \text{(A6)}$$

for some appropriate positive $\tilde{C}_r$. Noting that $t_j \leqslant t_{J^*} = O\{N^{-\delta/2}\surd(\log N)\}$ uniformly on $\mathscr{J}_N$, we may easily show by direct comparison that $N^{-2(1-1/r)/(v+vt_j^2)} = O(N^{-2(1-1/r)/v})$ as $N \to \infty$. Upon choosing $r = 2(v\delta)^{-1}$, we obtain the final bound for (A6) as $\tilde{C}_\delta \sup_z \sigma^4(z) N^{-2/v+\delta}$.

Case 2. Here $\sigma^2(i/N) \neq$ const, for $i = 2^{J-j}(k-1)+1, \ldots, 2^{J-j}k$, so that possibly $\mu_{1,j,k} \neq \mu_{2,j,k}$.

$$E\left\{(d_{1,j,k} - d_{2,j,k}) I\left(\frac{|d_{1,j,k} - d_{2,j,k}|}{d_{1,j,k} + d_{2,j,k}} > t_j\right) - (\mu_{1,j,k} - \mu_{2,j,k})\right\}^2$$

$$\leqslant 2E\left[\{d_{1,j,k} - d_{2,j,k} - (\mu_{1,j,k} - \mu_{2,j,k})\} I\left(\frac{|d_{1,j,k} - d_{2,j,k}|}{d_{1,j,k} + d_{2,j,k}} > t_j\right)\right]^2 + 2(\mu_{1,j,k} - \mu_{2,j,k})^2\,\text{pr}\left(\frac{|d_{1,j,k} - d_{2,j,k}|}{d_{1,j,k} + d_{2,j,k}} < t_j\right)$$

$$\leqslant 2w_{j,k}^2 + 2(\mu_{1,j,k} - \mu_{2,j,k})^2\,\text{pr}\left(\frac{|d_{1,j,k} - d_{2,j,k}|}{d_{1,j,k} + d_{2,j,k}} < t_j\right). \quad \text{(A7)}$$

If $\mu_{1,j,k} = \mu_{2,j,k}$ then the second summand disappears. Assume, without loss of generality, that $\mu_{1,j,k} > \mu_{2,j,k}$. Noting that $w^2_{i,j,k} \leqslant v \sup_z \sigma^4(z)/2$, we bound (A7) from above by

$$
\begin{aligned}
2v \sup_z \sigma^4(z) + 2(\mu_{1,j,k} - \mu_{2,j,k})^2 \,\mathrm{pr}\,\{&(d_{1,j,k} - \mu_{1,j,k})(t_j - 1) + (d_{2,j,k} - \mu_{2,j,k})(t_j + 1) \\
&+ 2\mu_{1,j,k} t_j > (1 + t_j)(\mu_{1,j,k} - \mu_{2,j,k})\} \\
&\leqslant 2v \sup_z \sigma^4(z) + \frac{2}{(1+t_j)^2}\{(1-t_j)^2 w^2_{1,j,k} + (1 + t_j^2) w^2_{2,j,k} + 4\mu^2_{1,j,k} t_j^2\} \\
&\leqslant 4v \sup_z \sigma^4(z) + 8 \sup_z \sigma^4(z) \log N = 4 \sup_z \sigma^4(z)(v + 2 \log N);
\end{aligned}
$$

the first inequality is a consequence of Markov's inequality. We now move on to the last step of the proof, the evaluation of the full L_2 risk. Define the set

$$
\mathcal{N}_N = \{(j, k) : \sigma^2(i/N) := \mathrm{const},\ i = 2^{J-j}(k-1) + 1, \ldots, 2^{J-j}k\};
$$

see Case 1 above. Denote by M the number of jumps in $\sigma^2(z)$. At each scale j, at most M indices (j, k) are in $\mathcal{N}^c_N$. We have

$$
\begin{aligned}
\frac{1}{N} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} &E(\mu_{j,k} - \hat\mu^{(h)}_{j,k})^2 \\
&= \frac{1}{N} \sum_{(j,k) \in \mathcal{I}_N \cap \mathcal{N}_N} E(\mu_{j,k} - \hat\mu^{(h)}_{j,k})^2 + \frac{1}{N} \sum_{(j,k) \in \mathcal{I}_N \cap \mathcal{N}^c_N} E(\mu_{j,k} - \hat\mu^{(h)}_{j,k})^2 + \frac{1}{N} \sum_{(j,k) \in \mathcal{I}^c_N} \mu^2_{j,k} \\
&\leqslant \frac{1}{N} \sum_{(j,k) \in \mathcal{I}_N \cap \mathcal{N}_N} \tilde C_\delta \sup_z \sigma^4(z) N^{-2/v+\delta} + \frac{MJ^*}{N} 4 \sup_z \sigma^4(z)(v + 2 \log N) \\
&\quad + \frac{M}{N} \sum_{j=J^*}^{J-1} \frac{2^{J-j}}{4} \left\{\sup_z \sigma^2(z) - \inf_z \sigma^2(z)\right\}^2 \\
&\leqslant N^{-1} 2^{J^*} \tilde C_\delta \sup_z \sigma^4(z) N^{-2/v+\delta} + \frac{MJ^*}{N} 4 \sup_z \sigma^4(z)(v + 2 \log N) \\
&\quad + \frac{M}{2} \left\{\sup_z \sigma^2(z) - \inf_z \sigma^2(z)\right\}^2 \frac{2^{J-J^*} - 1}{N} \\
&= O(N^{-2/v}) + O(N^{-1} \log^2 N) + O(N^{-(1-\delta)}) = O(N^{-\min(1-\delta, 2/v)}),
\end{aligned}
$$

as $N \to \infty$. This completes the proof of Theorem 1 for the case $(e) = (h)$. The proof for the case $(e) = (s)$ proceeds by showing that $\hat\mu^{(s)}_{j,k}$ and $\hat\mu^{(h)}_{j,k}$ are close in the mean-square sense, and then using the bounds for $\hat\mu^{(h)}_{j,k}$ obtained above. Full details are included in the technical report by the authors, available from the web address given in § 2·2. □

References

Antoniadis, A. & Fryzlewicz, P. (2006). Parametric modelling of thresholds across scales in wavelet regression. *Biometrika* **93**, 465–71.

Bera, A. K. & Higgins, M. L. (1993). ARCH models: properties, estimation and testing. *J. Econ. Surveys* **7**, 305–66.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Economet.* **31**, 307–27.

Clémençon, S. & Slim, S. (2004). Statistical analysis of financial time series under the assumption of local stationarity. *Quant. Finance* **4**, 208–20.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **16**, 1–37.

Dahlhaus, R. & Subba Rao, S. (2006). Statistical inference for time-varying ARCH processes. *Ann. Statist.* **34**. To appear.

Donoho, D., Mallat, S., von Sachs, R. & Samuelides, Y. (2003). Locally stationary covariance and signal estimation with macrotiles. *IEEE Trans. Sig. Proces.* **53**, 614–27.

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.
FAN, J. & YAO, Q. (2003). *Nonlinear Time Series*. New York: Springer-Verlag.
FISZ, M. (1955). The limiting distribution of a function of two independent random variables and its statistical application. *Colloq. Math.* **3**, 128–46.
FRYZLEWICZ, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Statist.* **32**, 503–28.
FRYZLEWICZ, P. & NASON, G. P. (2004). A Haar-Fisz algorithm for Poisson intensity estimation. *J. Comp. Graph. Statist.* **13**, 621–38.
GAO, H.-Y. (1997). Choice of thresholds for wavelet shrinkage estimate of the spectrum. *J. Time Ser. Anal.* **18**, 231–52.
GIRAITIS, L., LEIPUS, R. & SURGAILIS, D. (2005). Recent advances in ARCH modelling. In *Long Memory in Economics*, Ed. A. Kirman and G. Teyssiere, pp. 3–39. Berlin: Springer.
HULL, J. C. (1997). *Options, Futures, and Other Derivatives*, 3rd ed. London: Prentice-Hall.
JOHNSON, N. L., KOTZ, S. & BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*, 2nd ed. New York: Wiley.
KITAGAWA, G. & AKAIKE, H. (1978). A procedure for the modeling of non-stationary time series. *Ann. Inst. Statist. Math.* **30**, 351–63.
KOKOSZKA, P. & LEIPUS, R. (2000). Change-point estimation in ARCH models. *Bernoulli* **6**, 513–39.
MERCURIO, D. & SPOKOINY, V. (2004). Statistical inference for time-inhomogeneous volatility models. *Ann. Statist.* **32**, 577–602.
MIKOSCH, T. & STĂRICĂ, C. (2004). Non-stationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev. Econ. Statist.* **86**, 378–90.
MOULIN, P. (1994). Wavelet thresholding techniques for power spectrum estimation. *IEEE Trans. Sig. Process.* **42**, 3126–36.
NASON, G. P. & SILVERMAN, B. W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics*, Ed. A. Antoniadis and G. Oppenheim, Lecture Notes in Statistics, **103**, pp. 281–300. New York: Springer-Verlag.
NASON, G. P., VON SACHS, R. & KROISANDT, G. (2000). Wavelet processes and adaptive estimation of the evolutionary spectrum. *J. R. Statist. Soc.* B **62**, 271–92.
NEUMANN, M. & VON SACHS, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In *Wavelets and Statistics*, Ed. A. Antoniadis and G. Oppenheim, Lecture Notes in Statistics, **103**, pp. 201–29. New York: Springer-Verlag.
OZAKI, T. & TONG, H. (1975). On the fitting of non-stationary autoregressive models in time series analysis. In *Proceedings of the 8th Hawaii International Conference on System Sciences*, pp. 224–6. Hawaii: Western Periodical Co.
RUDZKIS, R., SAULIS, L. & STATULEVICIUS, V. (1978). A general lemma on probabilities of large deviations. *Lithuanian Math. J.* **18**, 226–38.
SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons.
STĂRICĂ, C. & GRANGER, C. W. J. (2005). Non-stationarities in stock returns. *Rev. Econ. Statist.* **87**, 503–22.
TAYLOR, S. J. (1986). *Modelling Financial Time Series*. Chichester: John Wiley & Sons.
TONG, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.
VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.
WAHBA, G. (1980). Automatic smoothing of the log-periodogram. *J. Am. Statist. Assoc.* **75**, 122–32.
WICKERHAUSER, M. V. (1994). *Adapted Wavelet Analysis from Theory to Software*. Wellesley, MA: A. K. Peters, Ltd.

[*Received June* 2005. *Revised January* 2006]