WAVELET TRANSFORMS FOR THE STATISTICAL ANALYSIS OF RETURNS GENERATING STOCHASTIC PROCESSES

ENRICO CAPOBIANCO*

CWI, Kruislaan 413, 1098 SJ Amsterdam (NL)

We study high frequency Nikkei stock index series and investigate what certain wavelet transforms suggest in terms of volatility features underlying the observed returns process. Several wavelet transforms are applied for exploratory data analysis. One of the scopes is to use wavelets as a pre-processing smoothing tool so to de-noise the data; we believe that this procedure may help in identifying, estimating and predicting the latent volatility. Evidence is shown on how a non-parametric statistical procedure such as wavelets may be useful for improving the generalization power of GARCH models when applied to de-noised returns.

Keywords: Financial volatility, multiresolution analysis, wavelet transforms, data de-noising with wavelet shrinkage, GARCH models

1. Introduction

Volatility models have characterized the field of financial time series in the last two decades; they work through equations for the conditional variance together with the conditional mean and aim to improve the quality of predictions for stock returns, rates and many other derivatives. The idea is to better exploit the available information, since this comes with the data under some particular form of dependency, precisely the one in the conditional variance. At the same time, the goal of providing effective models for *financial returns generating stochastic processes* has found an original solution with the *autoregressive conditional heterescedasticity* (ARCH) models [8] and their generalized representation (GARCH) [2]; many other related characterizations for this kind of stochastic processes followed afterwards.

Since then, new mathematical and statistical tools have become available for applications in financial time series [14]. From an empirical viewpoint there are also several new perspectives; for instance, in line with the direction of studies [12] which aim to see financial markets as a place where people act according to different time horizons for what concerns their investment decisions, one may try to interpret

---

*E-mail: capobianco@mail.webmedia.it

the observed returns as a signal that can be examined at different resolution levels, which reflect the investors' time horizons. We also consider the fact that in general a signal might seem stationary at first observation, but at a more detailed level of analysis discontinuities can appear so that previously undetected non-stationarity behaviour could now show up. This aspect might be relevant in financial time series analysis, where wavelets could be useful for dealing with the task of exploring latent data features; their inherent multiresolution property may help in explaining the time/space and frequency varying components in a signal. It is important to find new flexible tools for modelling non-stationary stochastic processes, especially when it is important to emphasize the contribution that local features of the observed signals can offer in order to find information relevant for forecasting purposes.

In Sec. 2 we briefly introduce wavelets. In Sec. 3 we describe the data set that we use for the analysis, together with the wavelet families we adopted. In Sec. 4 we present some results of the exploratory analysis performed on the observed time series of daily Nikkei index returns. In Sec. 5 we look at various wavelet decomposition techniques and show their reconstruction power. Data de-noising is the topic of discussion in Sec. 6, and an interesting example of application of a wavelet estimator is described in Sec. 7. In Sec. 8 we report the conclusions.

2. Wavelets and their Properties: A Brief Review

Financial time series data sets are temporal series inherently perturbed by noise; stock market prices can be affected by so many factors and by so many different institutions and individuals that no theory can suggest a safe way of modelling data to reflect price movements. Therefore for financial time series there exists no tool for extracting the true signal and thus separating it from the noise in the observed values; of course one can hope to build a model which is able to approximate the sought for signal, but how well this happens through the part of price variation that the designed model manages to explain is simply not known because of the presence of noise and because of the same limitations and contraints imposed by the model.

We think that the ability to separate the true underlying volatility-carrier signal from the pure noise might potentially be improved if we could look deeply into the data, i.e. by analyzing the signal at different resolution levels. One goal is to find the best strategy for decomposing the signal through a wavelet expansion, with basis functions able to capture the main characteristics of the time series and suitable to be interpreted. The multi-resolution view of a signal is the strength of the wavelet transform; with a simple prototype function we can perform a fine spatial/temporal-and-frequency analysis through a contracted (high frequency) and a dilated (low frequency) version of the same function respectively. The wavelet transform is useful from this last perspective, more than other techniques, since it gives a resolution which is sharper in time(space)/frequency at respectively high/low frequencies and therefore offers more flexibility from its localization power.

Consider a general function f which we want to expand in terms of some basis functions with certain time-frequency localization properties; given the **scaling**

function (or **father wavelet**) ϕ such that its dilates and translates constitute orthonormal bases for all the V_j subspaces that are scaled versions of the subspace V_0 to which ϕ belongs, we can form a **Multiresolution Approximation** (MRA) of $L^2(R)$ once some properties are satisfied (see [6, 11] for technical definitions and details).

With a DWT (i.e. a **Discrete Wavelet Transform**) we are basically constructing a map $f \rightarrow w$ from the signal domain to the wavelet coefficient domain, or in other words we apply the transformation $w = Wf$. Consider now a **mother wavelet** ψ and its derived terms indicated with ψ_{jk} (j is the dilation or level index and k is the translation or shift index), which are obtained as:

$$\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k) \tag{1}$$

For certain ψ's, the ψ_{jk} form an orthonormal basis for functions in some particular spaces. A general wavelet decomposition is described by $f(x) = \sum_{jk} f_{jk}\psi_{jk}(x)$, where f_{jk} give the information about the function f near time point $2^j k$ and near frequency proportional to 2^j. The ϕ and ψ pair of functions generate the series of approximating spaces V_j addressed above. At a more specific level of analysis, the DWT algorithm is able to produce coefficients for fine scales, thus capturing high frequency information, and for coarse scales, thus capturing low frequency information. Therefore, a sequence of smoothed data and a sequence of details not previously accounted for that give information at finer resolution levels, are obtained. We come up with a representation like:

$$f(x) = \sum_k c_{j0,k}\phi_{j0,k}(x) + \sum_{j>j0}\sum_k d_{j,k}\psi_{j,k}(x), \tag{2}$$

where $\phi_{j0,k}$ is a scaling function with the corresponding coarse scale coefficients $c_{j0,k}$ and $d_{j,k}$ are the detail (fine scale) coefficients; the first term of the right hand side of (2) is the projection of f onto the coarse approximating space V_{j0} while the second term represents the detail. We can define $c_{j,k} = \frac{1}{n}\sum_{i=1}^n \phi_{j,k}(x_i)$ and $d_{j,k} = \frac{1}{n}\sum_{i=1}^n \psi_{j,k}(x_i)$. A clear advantage of an orthogonal wavelet expansion is the resulting independence among coefficients; this mapping from the signal to the wavelet coefficients domains allows one to perform statistical inference in the projected domain.

3. The Data Set and the Wavelet Family for the Analysis

The data set we analyze is the daily Nikkei index, with observations ranging from May 17, 1949 to 31 July 1996, for a total amount of 13505 data. We construct the series of returns in the usual fashion, i.e. $r_t = \ln(p_t/p_{t-1}) \times 100$.[a] The wavelet family

---

[a]There are clearly few big outliers due to well-known shocks occurred worldwide. We do not eliminate them from the sample at hand, for the reasons that first it is always an hard task with time series but particularly because we want to leave their treatment to some de-noising procedure able to discriminate between pure signal and disturbance. We observed that by using a *robust-cleaner wavelet smoother*, we ended up losing too much structure in the sample and thus we avoid pre-processing the data in this way.

chosen among the many available is the **symmlet**; for this data set symmlets and **coiflets**, another family with similar features, behave similarly. Symmlets-8 have a compact support, are orthogonal, nearly symmetric and with good smoothness properties [6]; therefore they are well localized in time, thus delivering good *spatial adaptivity*, and where a smooth signal is found they can be represented with relatively few coefficients, i.e. satisfy the *principle of sparsity of representation.*

With wavelets we basically adopt a flexible degree of smoothing according to the resolution level. Thus, by increasing the resolution level j we decrease smoothing and vice versa when we decrease j, just as in the case of using a variable bandwidth for each time location in a kernel smoother. Other considerations have to be done with regard to the choice of working with decimated wavelets instead of (ST) stationary ones, i.e. non-decimated. The main consequence is the number of coefficients retained in the analysis at each resolution level investigated; in the first case every time we switch from one level to another we have half the number of coefficients available compared to the number used in the previous higher resolution level, while for stationary wavelets this decimation does not occur. The reason behind the importance of using all the coefficients in some applications, comes from the fact that a more precise alignment with data features can be found at every resolution level. The price to pay is that many more coefficients remain in the analysis.

Models for time series or econometric analysis should be built according to the principle of *parsimony*, which means to use as bare a structure (i.e. parameters) as possible. In wavelets the same aspect is brought in by the concept of *sparsity*, which means to be able to approximate a function belonging to a certain space by projecting it onto a sequence of sub-spaces at different resolution levels and using relatively few coefficients in the function representation. Thus, the advantage of a sparse representation is obtained when many components of the coefficient vector w can be considered negligible for reconstructive power purposes, and therefore eliminated. However, two problems are encountered with financial time series models: (a) the non-parametric nature of the wavelet transform does not offer the usual interpretation for the estimated coefficients as expected by financial econometricians (but by no means should this aspect prevent them from being studied) (b) care must be taken in analyzing the coefficients selected due to the highly noisy nature of these data; one thus needs reliable procedures to get rid of coefficients considered not useful for the signal reconstruction.

4. Exploratory Analysis of Nikkei Index Data

The notation we adopt here follows the one used in [3]. Thus, with $s_{j,k}$ and $d_{j,k}$ we indicate the smooth and detail coefficients that appear in the signal decomposition. Generally speaking, with decimated wavelets and the sample size n divisible by 2^J we have $\frac{n}{2}$ coefficients $d_{1,k}$, i.e. the finest scale, $\frac{n}{4}$ $d_{2,k}$ (the next finest scale) and so forth until we find $\frac{n}{2^J}$ $d_{J,k}$ (the coarsest

Table 1.   Energy percentages by resolution levels.

resolution levels	d1	d2	d3
coiflet-6	0.435	0.282	0.138
symmlet-6	0.459	0.262	0.138
symmlet-8	0.455	0.257	0.142

Table 2.   Energy percentages by number of coefficients.

n. coefficients	1	136	271	406	876	1351	2026	3377
coiflet-6	0.014	0.262	0.359	0.427	0.523	0.668	0.756	0.860
symmlet-6	0.016	0.262	0.357	0.423	0.519	0.662	0.750	0.856
symmlet-8	0.016	0.268	0.363	0.430	0.527	0.665	0.751	0.856

scale), for a total amount of $n = \frac{n}{2} + \frac{n}{4} + \cdots + \frac{n}{n^J}$ coefficients.[b] The detail coefficients embed information about finer and finer resolution levels, thus offering the true advantage of wavelets compared to other smoothing techniques. We can represent the wavelet coefficients as $w = [s_J, d_J, \ldots, d_1]'$, where $s_J = [s_{J,1}, s_{J,2}, \ldots, s_{J,\frac{n}{2^J}}]'$, $\ldots, \ldots, \ldots, d_1 = [d_{1,1}, d_{1,2}, \ldots, d_{1,\frac{n}{2}}]'$. The original signal can be decomposed according to $S_{j,k}(t) = s_{j,k}\phi_{j,k}(t)$, $D_{j,k}(t) = d_{j,k}\psi_{j,k}(t)$, so that we have $S_j(t) = \sum_k S_{j,k}(t)$ and $D_j(t) = \sum_k D_{j,k}(t)$, and the the signal can be represented as $f(t) = S_J(t) + D_J(t) + D_{J-1}(t) + \cdots + D_1(t)$. This is an MRA of the signal and the goal now is to operate a selective *Multi-resolution Decomposition* of it by extracting the most informative components.

Table 1 gives the percentages of energy distributed in the three most relevant resolution levels for decimated wavelets; Table 2 gives instead the percentages of energy distributed according to groups of coefficients ordered by decreasing size. The total energy (see [5]) is given by $E = \sum_{k=1}^{n} f_k^2 = E_J^s + \sum_{j=1}^{J} E_j^d$, where $E_J^s = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^J}} s_{J,k}^2$ and $E_j^d = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^j}} d_{j,k}^2$, $j = 1, \ldots, J$.

One can see that the choice between one or another family, at least according to a visual inspection and a comparison of energy percentages, doesn't represent an issue indeed, provided that a certain smoothness is allowed (here the smoothness index is the number attached to the wavelet name). We chose to work with symmlet-8 (Fig. 1) and when we reconstruct the original signal with the Inverse DWT function, we may notice how the reconstruction improves by going from level d_3 to d_1 (Fig. 2).

The energy enclosed by coefficients in the remaining resolution levels does not turn out to be so relevant, also at a visual inspection, even if they still enclose

---

[b]The number of coefficients at the various levels is not exact if n is not divisible by 2^J, but only approximate, but still the total number of coefficients is n, even if not exactly $\frac{n}{2^j}$ at scale 2^j.
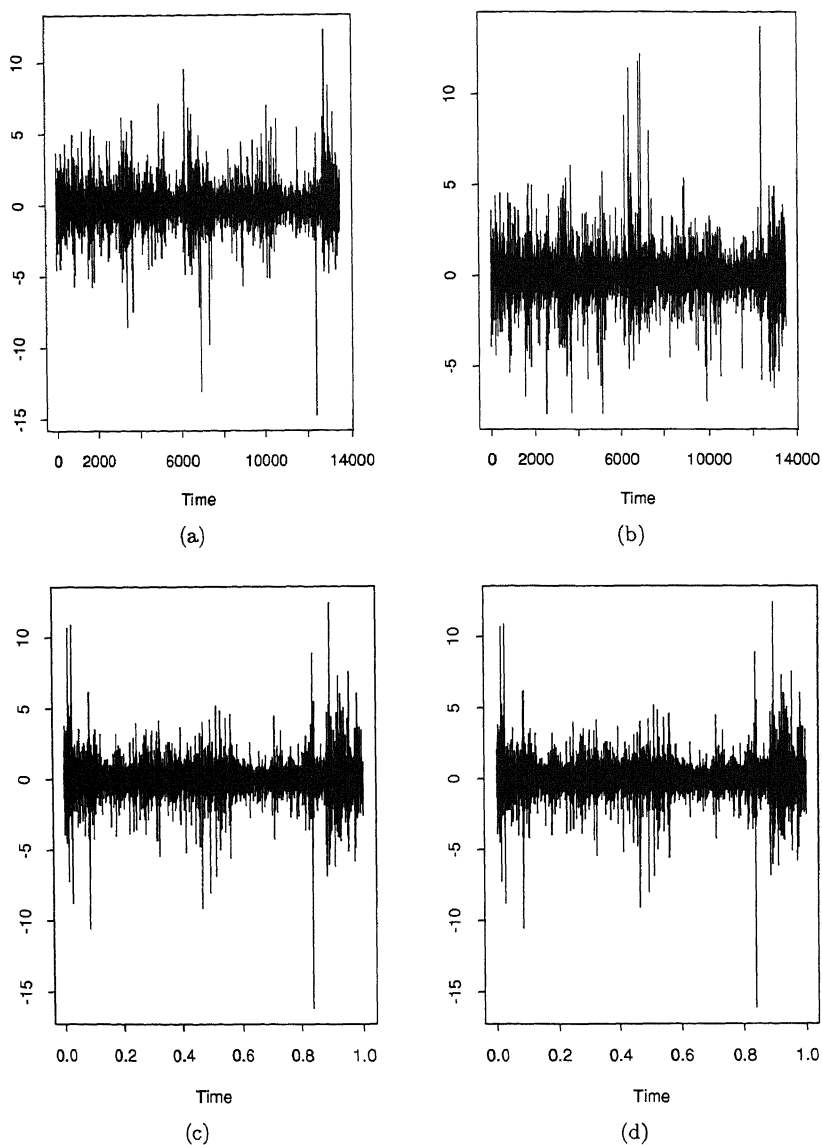
Fig. 1.   (a) symmlet-8 transform. (b) coiflet-6 transform. (c) original signal. (d) signal reconstructed from symmlet-8.

residual energy. With ST wavelets, where decimation does not occur, n coefficients appear at every resolution level. We found the following percentages of energy distributed among levels for a ST symmlet-8: $d_1 = 0.109$; $d_2 = 0.133$; $d_3 = 0,133$; $d_4 = 0.137$; $d_5 = 0.132$; $d_6 = 0.169$; $s_6 = 0.188$.

At first it seems that such a homogeneous distribution of energy does not help too much in separating the components relevant for the reconstruction. But we
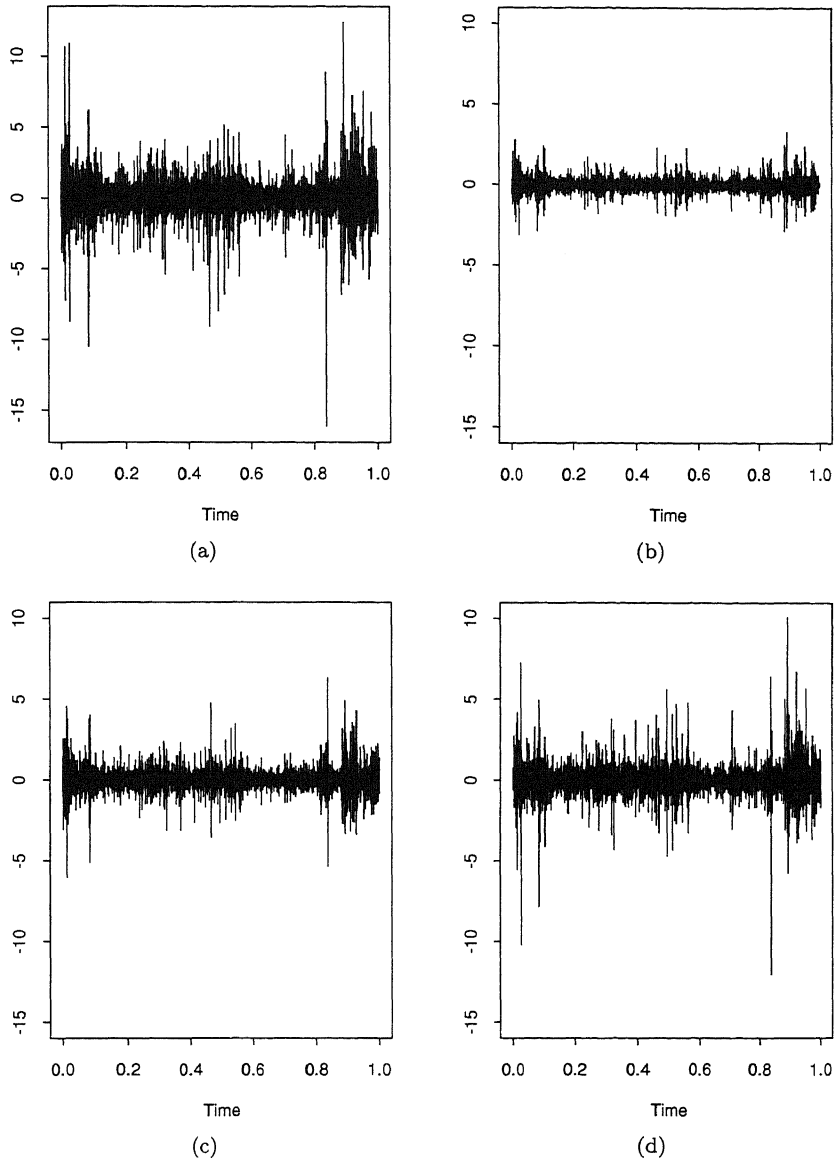
Fig. 2. Reconstruction from d3 ((a),(b)), d2 ((c),(d)) and d1 levels.

also have to consider the result in the light of the nature of the signal at hand, i.e. a financial time series, where we expect to find out short-, mid- and long-term information related to the various market horizons of different agents operating in the market. From Fig. 3 we can notice a better resolution power at almost every level, but the redundancy of coefficients that by default comes with ST wavelets is also relevant from the perspective of selectively reconstructing the signal (i.e. by
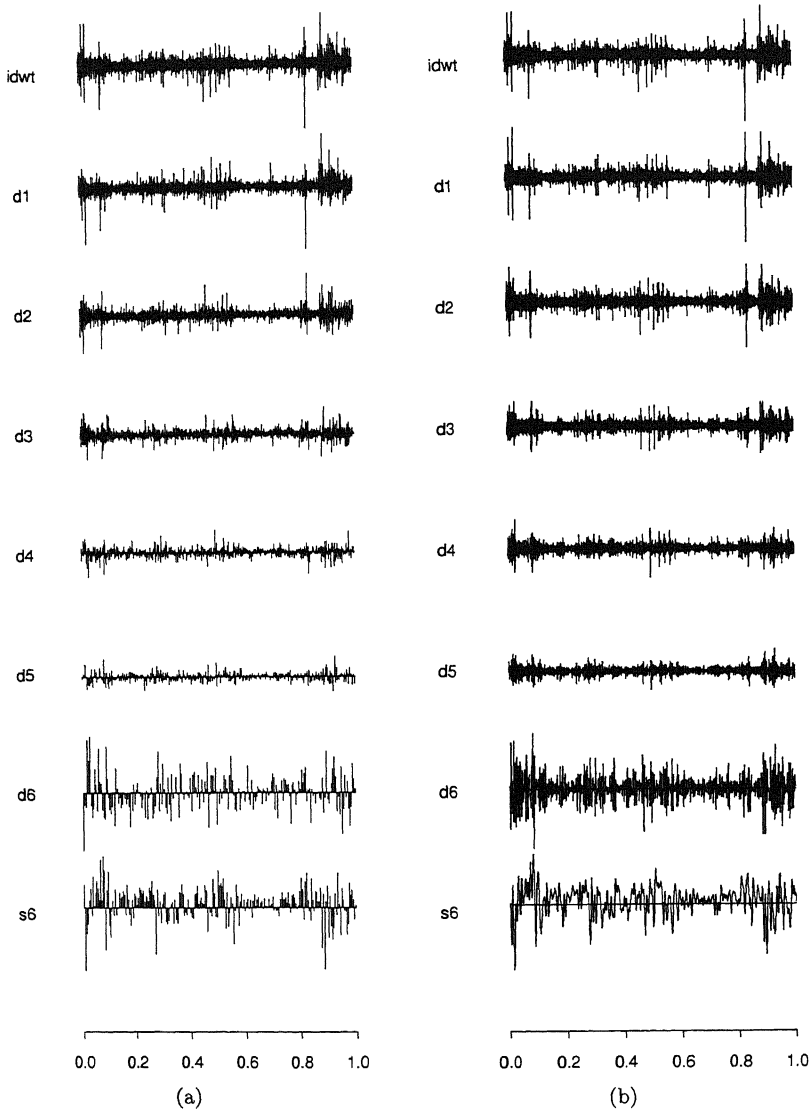
Fig. 3.   Multiscale decomposition with decimated (left) and ST symmlet-8.

levels of resolution) when compared to the decimated wavelets shown before.[c] The relevant numbers of coefficients present in the decimated symmlet-8 are 6752 in d_1, 3376 in d_2 and 1688 in d_3; for the stationary wavelets the corresponding values are of course the same at every level.

[c]The length of the lines is relative to the magnitude of the coefficients for each level according to a particular scale and the coefficients are spaced so to observe their localization properties, which helps in explaining where in the function significant changes occur.

A final observation is made: structure mixed with noise to different degrees is discovered in different resolution levels in both cases, i.e. with decimated or not decimated wavelets, thus suggesting the possibility that latent volatility features characterizing different scales may reflect different market horizons for operators.

5. Other Wavelet Decomposition Techniques

5.1. *Wavelet packets*

There are other procedures which are based on the wavelet transform and are useful for our analysis. **Wavelet packets**, for instance, allow for the presence of an oscillation parameter to consider periodic behaviour in the series; since we can combine wavelet functions so as to build **tables** or **dictionaries**, we obtain a better domain of wavelets, compared to the basic one, from which to select a basis that represents the signal. We can still select an orthogonal transform from the so-called **Wavelet Packet Table** (WPT), something perfectly equivalent to the DWT employed before. But we can do more indeed; we can try to choose the best basis according to the procedures suggested by [5]. In general, when we extract components from a WPT we obtain, following [3], a decomposition like:

$$W_{j,o}(t) = \sum_k w_{j,o,k} W_{j,o,k}(t) \tag{3}$$

where the W components play the same role of the D's before. We can select entire resolution levels (see Fig. 4) in order to test their individual reconstructing power and we can design special WPT from which to search the best basis representing the signal through a specific selection of sets of coefficients. A **crystal** is a set of wavelet coefficients, which for the WPT is indexed by the level j and the oscillation b: $W_{j,b} = (w_{j,b,1}, w_{j,b,2}, \ldots, w_{j,b,n/2^j})'$.

The level 1 crystals in Fig. 4 have scale 2 and correspond to the DWT coefficients previously indicated by s_1 and d_1; therefore, it brings signal information at the highest resolution level and it is ordered by increasing oscillation index. The decomposition is thus obtained, as from (3). The same applies for the level 2 crystals and decomposition.

5.2. *Cosine Packects*

With **cosine packets** we use instead cosine functions localized in time that form smooth basis functions. The **Discrete Cosine Transform** (DCT) is the discretized version of the **Fourier Cosine Transform** of a signal, i.e. $DCT_k = \sqrt{\frac{2}{n}} s_k \sum_{i=0}^{n-1} f_{i+1} \cos(\frac{(2i+1)k\pi}{2n})$, where $k = 0, 1, \ldots, n-1$ and s_k is the scaling factor equal to 1, if $k = 0$ or $k = n$, or to $1/\sqrt{2}$, if k is different from the previous values. An orthogonal transformation that maps a signal from the time to the frequency domain is thus obtained. For the DCT, depending on the **taper functions** we choose, we can design cosine packets which improve the time localization power and thus
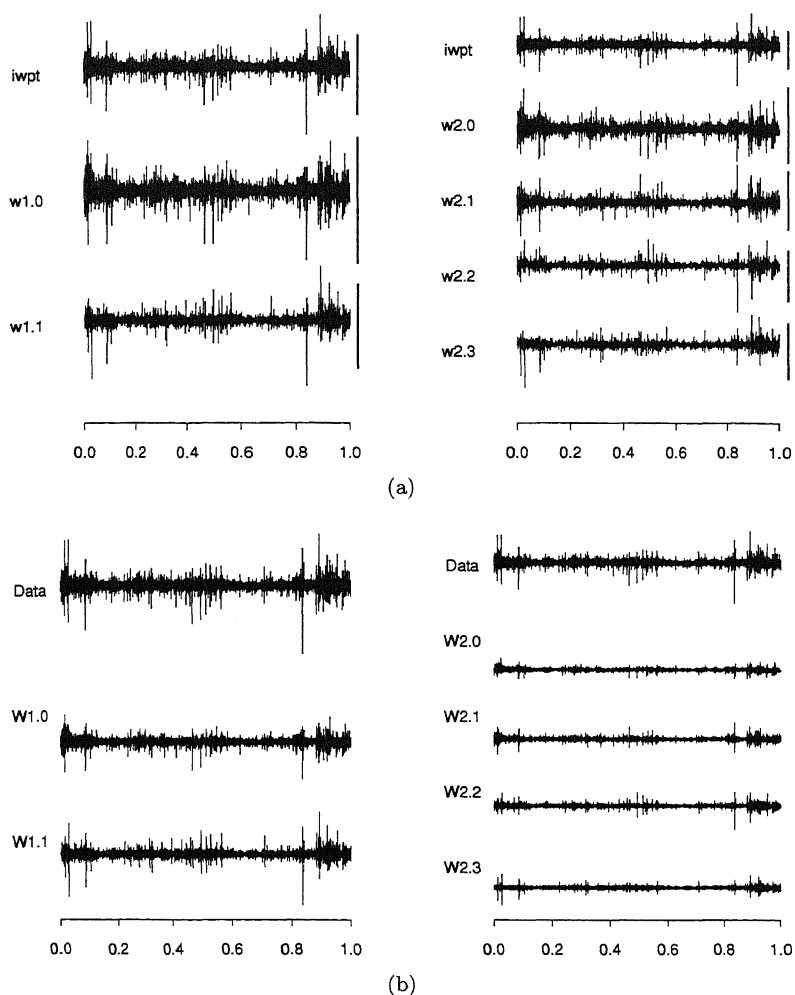
Fig. 4.   Reconstructed signal from w1 and w2 crystals (a) and from W1 and W2 signals.

capture local features in the data. One simply creates smooth basis functions by letting cosine functions go to 0 in a selected interval.

We have computed experiments with the **Cosine Packet Table** (CPT), and one was the application of the **best basis algorithm** of [5], i.e. a global optimization procedure for finding the transform that best matches the signal features. It is applied by searching for the minimum of the cost function $\sum_{j,o} E(W_{j,o})$, which is like searching for a *minimum entropy transform*. We have noted that in terms of reconstructing the signal we have pretty much the same power as we had before with the DWT. Thus, we analyzed another algorithm, presented in the following sub-section, which is more effective in dealing with non-stationary signals and capturing local features.

Table 3. Number of largest coefficients per type of transform and level with associated percentage of energy according to the MP algorithm decomposition.

levels	0	1	2	3	5
wp-transf.	19(0.799)	13(0.152)	8(0.038)	–	–
cp-transf.	18(0.731)	7(0.092)	–	4(0.039)	11(0.106)
dwt-transf.	23(0.713)	14(0.221)	8(0.053)	–	–

5.3. *Matching pursuit basis selection*

The Matching Pursuit (MP) algorithm of [9] decomposes a signal as a sum of *atomic waveforms* belonging to *dictionaries*, like WPT or CPT, and other function families too. The MP decomposition is not a global optimization procedure and does not obtain an orthogonal decomposition. It is a greedy algorithm which iteratively, at successive steps, decomposes the residual term left from a projection of the signal onto the elements of a selected dictionary in the direction of that atom which best matches the signal features. In summary, the algorithm approximates a function as $f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + \text{res}_i(t)$ by computing at each H_{γ_i} the quantity $\mu_{\gamma,i} = \int \text{res}_{i-1}(t) H_\gamma(t) dt$ and by finding $\gamma_i = \arg\min_{\gamma \in \Gamma} \|\text{res}_{i-1}(t) - \mu_{\gamma,i} H_\gamma(t)\|$. Then the updated residual is given by $\text{res}_i(t) = \text{res}_{i-1}(t) - h_i H_{\gamma_i}(t)$ and the procedure is repeated until $i \leq n$.

The results reported in Table 3 show the better localization at the high frequencies for MP on WPT compared to CPT and DWT. The reported values correspond to the largest coefficients found at each level after the decomposition, with the energy percentages appearing in parentheses. For the WPT the MP finds that 0.799 is the percentage of energy explained by the 19 largest coefficients, which is a better performance compared to the MP applied on CPT, which explains less, 0.731, with almost the same number of coefficients, 18, and compared to the equivalent DWT (obtained as a special case with a linearly independent wavelet packet transform), which needs 23 largest coefficients to explain less energy percentage, 0.713. These results are obtained for the best resolution available, by choosing coarser levels CPT spreads more information than WPT and DWT needs comparatively more coefficients, and thus is less sparse as a representation than the one obtained when MP runs on WPT.

In any case, results indicate that there is no great change in performance when switching from one transform to another, the reason being that the signal at hand is less characterized by periodicities than it is instead by time non-uniformities observed in the series. This fact has an high relevance when one tries to exploit the power of redundant but richer classes of functions, like WPT or CPT, like it has been shown in [4] with intradaily data. Here the two function approximations for WPT and CPT are given by $f(t) = \sum_{jok} w_{j,o,k} W_{j,o,k}(t) + \text{res}_i(t)$ and $f(t) = \sum_{jok} c_{j,o,k} C_{j,o,k}(t) + \text{res}_i(t)$.

6. Denoising with Wavelets

The **wavelet shrinkage principle** of [7] applies a thresholded de-noising procedure to the data by shrinking wavelets coefficients to zero so that a limited number of them will be considered for reconstructing the signal. The fact that the noise is removed from the signal to obtain a better reconstruction might be crucial for financial time series in order to capture the underlying volatility structure. From the perspective of statistical inference, we are clearly employing a non-parametric procedure given that it does not rely specifically on assumptions about the underlying nature of the function $f(t)$ and it adopts a criterion similar to a locally adaptive bandwidth. Thus, denoising with wavelets is useful for spatially heterogeneous signals like financial time series.

The following algorithm implements the wavelet shrinkage principle:

- *DWT is applied to the data to make the empirical wavelet smooth and detail coefficients*
- *the wavelet coefficients, in particular at the finest scales, are shrunken toward zero by thresholding*
- *the inverse DWT is applied to the thresholded coefficients to reconstruct the signal*

Figures 5 and 6 show the signal and residuals extracted through de-noising runs with decimated and undecimated symmlets respectively. One can observe that in the right upper parts of the two figures box plots of level-by-level wavelet coefficients are reported and the coefficients within the marked central bands are eliminated because they are no not different from noise, according to thresholding procedure we adopted. The signal-to-noise ratio, still on a level-by-level basis, reported in the right bottom part of the figures, reflects a different number of wavelet coefficients used at each resolution level with the two different transforms.

The results indicate that decimated and ST wavelets are able to discriminate between signal and residuals quite clearly and in a different way according to the resolution level considered. A *level-adaptive threshold* is chosen for both the cases because from the experiments done it is the method that allows for a better signal/noise separation. It works according to the *soft shrinkage* rule selected as:

$$\delta_c(x) = \text{sign}(x)(|x| - c) \qquad (4)$$

when $|x| > c$; otherwise $\delta_c(x) = 0$. The threshold which adapts to each resolution level is based on the principle of minimizing at each resolution level the *Stein Unbiased Risk Estimator*, or *SURE*, such that the resulting estimator is quoted in the literature as **SureShrink**. It takes the following form $\lambda_j = argmin_{t \geq 0} SURE(d_j, t)$, with

$$SURE(d_j, t) = K - 2 \sum_{k=1}^{K} I_{[|d_{j,k}| \leq t\sigma_j]} + \sum_{k=1}^{K} \min \left[\left(\frac{d_{j,k}}{\sigma_j} \right)^2, t^2 \right] \qquad (5)$$
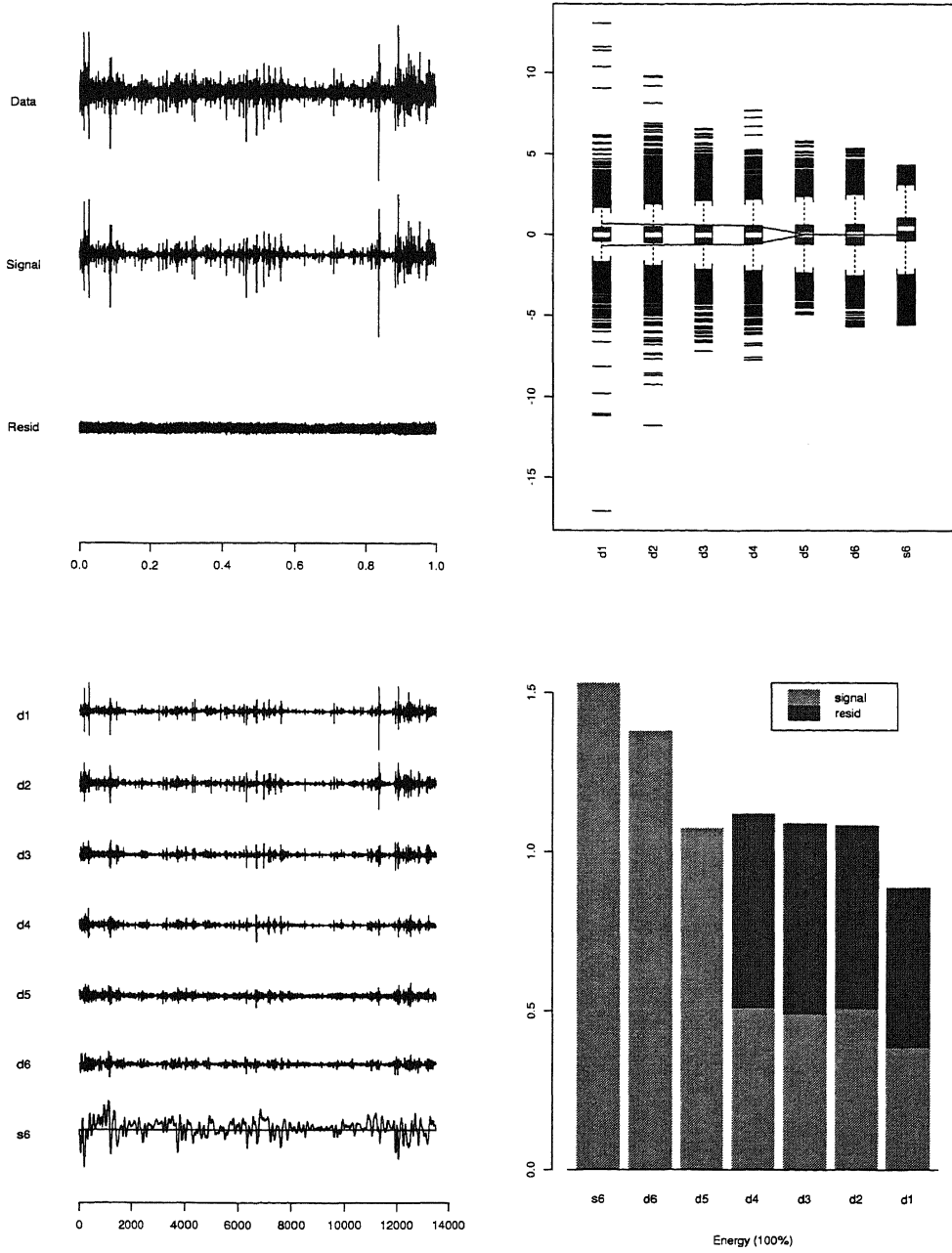
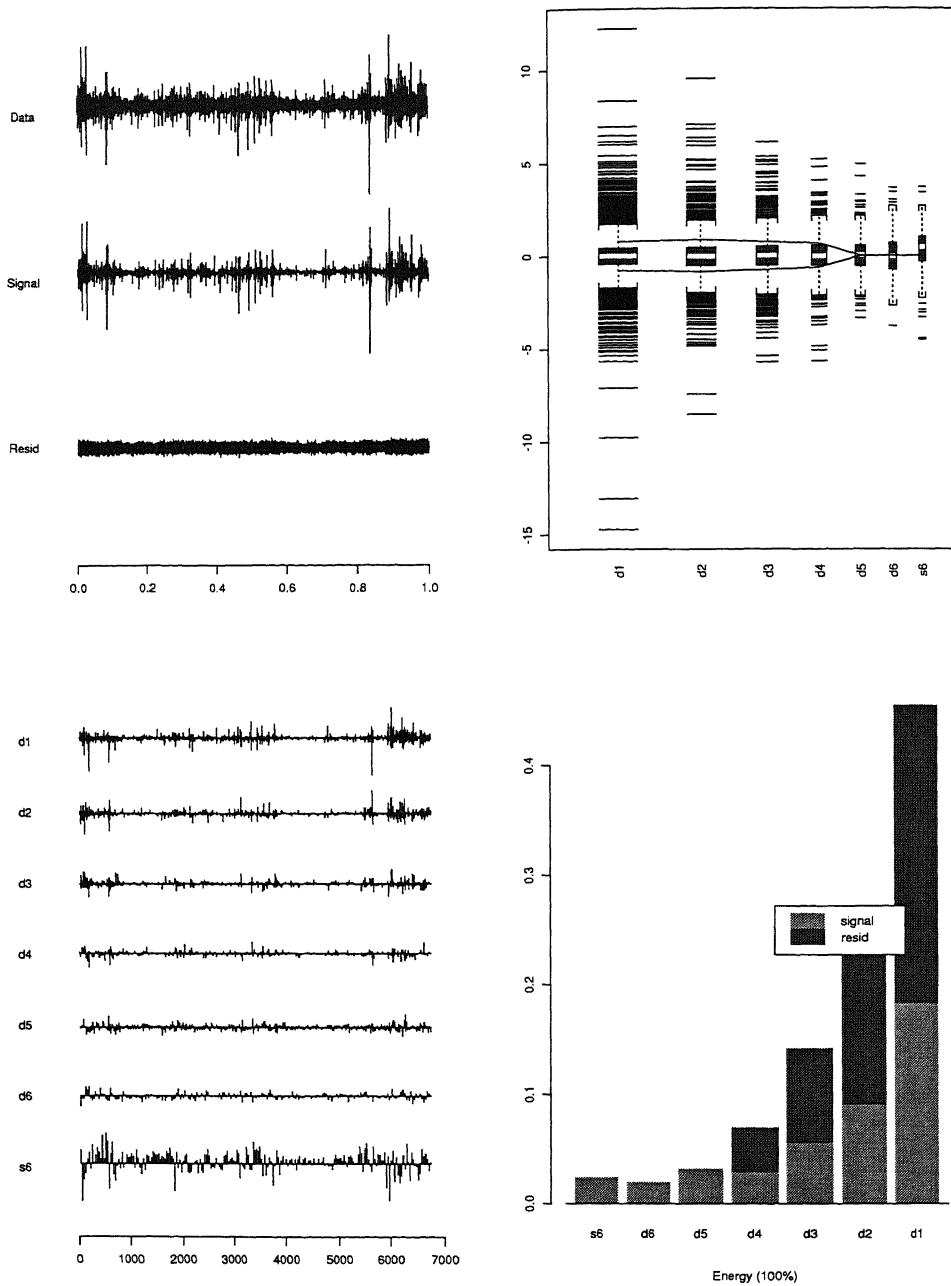Fig. 5.   Denoising with ST symmlets (adaptive threshold).

Fig. 6.   Denoising with decimated symmlets (adaptive threshold).

and where the shrinkage function depends also on the estimate of the scale of the noise; we found that performances are pretty much similar when different estimated scale functions are tried out.

Residual diagnostics[d] show that the autocorrelation functions and the residuals quantile plots for detecting deviations from the Gaussian distribution, suggest that model misspecification is still present; this is a real dilemma with unknown solution, and there is few things that either parametric or nonparametric statistical inference can do. However, one can choose to learn more about what is behind the data, like for instance to look at the volatility structure as we chose to do, since conditional mean misspecification may not be so unbearable as to prevent a reliable investigation of it and still allow for consistent conditional variance predictions to be achieved, as shown by [13].

Reducing the number of wavelet coeffcients is a problem that one has to deal with when an effort is made to build up a wavelet-based model. With regard to the de-noising procedures we adopted, there is still an high number of decimated wavelet coefficients left for reconstructing the signal; at the d1 level, for instance, signal coefficients in the various experiments (i.e. with different parameters in SURE) range from 1720 to 1751, in level d2 from 729 to 1061 and in level d3 from 460 to 822, while for ST wavelets (for one choice of parameters in SURE) the correspondent values are 4219, 5216 and 5988, with a number equal to n for coarsest scales (the ones not affected by the shrinkage algorithm because considered less noisy).

This method leads naturally to a selection of coefficients that embed the signal features of the data at hand, thus allowing for an excellent reconstruction. One could also select the resolution levels to which the coefficients belong, and try to pick out the specialized information brought at every scale. Then, one possibility is to use the information brought by these coefficients as regression coefficients in a model with a regression matrix composed of wavelet dilations and translations. It would be even better to allow these coefficients to be dynamically changing, in the style for instance of state space models, where we could set them into the state vector subject to recursive estimation by a Kalman Filter type of algorithm. The framework we have introduced here is in other words open to further refinement toward the idea of making wavelets more leading to build a model for representing observed data and related dynamics and thus for approximating signals, rather than simply pre-processing data, even if in a very informative way.

7. An Application of the SureShrink Estimator Combined with GARCH Modelling

We build our model with a simple framework, that of a *signal + noise* model, i.e. $y_t = f_t + \epsilon_t$, and apply it to the observed returns: the wavelet transform. In statistical terms one would say that the given model represents a semi-parametric

---

[d]Available from the author.

regression fit, where the f_t signal that we want to detect is assumed to behave as an ARCH process, i.e. $f_t = \xi_t \sigma_t$, with $\sigma_t = \sqrt{h_t}$ and $h_t = \sum_i b_i f_{t-i}^2$, as in the case for the ARCH natural generalization, the GARCH process, where lagged volatility values are included. In this way, we are basically emphasizing two facts: (i) our true signal component is inherently noisy and (ii) we want to limit the influence of noise on our data features. The non-parametric regression model derives from superposing the signal+noise model to a signal with GARCH-type disturbances; these disturbances will be assumed to follow a *Student's t* distribution, for dealing with the leptokurtosis of returns, while the additive noise is instead a general i.i.d. process. The model can be described as follows:

$$y_t = f_t + \epsilon_t \tag{6}$$

$$f_t = \xi_t \sigma_t, \quad \text{with} \quad \sigma_t = \sqrt{h_t} \tag{7}$$

$$h_t = \sum_i a_i h_{t-i} + \sum_i b_j f_{t-j}^2 \tag{8}$$

where $\epsilon_t \sim i.i.d.(0, \sigma_\epsilon^2)$, $f_t | \Psi_{t-1} \sim i.i.d.(0, h_t)$ and $\xi_t \sim N(0,1)$, and given the set of past information Ψ_{t-1}.

The semi-parametric form of the model thus depends on the choice of leaving unknown the conditional mean noise distribution, while selecting a specific parametric family for the disturbance affecting the GARCH-type signal.

This model, which we estimate in a sequential, i.e. two-step, fashion, may be more soundly justified when higher frequency or tick-by-tick data sets are used, perhaps, since in these contexts financial theory finds the presence of noise at a microstructural level. Nevertheless, we think it is interesting to analyze daily data because in this case the measurement error is relevant for the accuracy of predicting the volatility function. We aim to confirm that by de-noising the data. We got similar results just as if we had increased the frequency of observation, and thus could obtain better volatility predictions, confirming what has been shown in other studies [1]. We do not claim that one gets a better fit without the noise in the data, simply because the data to which the models apply are different; but we believe it is a legitimate argument to look at which of the two specified models turn out to be more informative for our purposes. Thus, we try to let a different signal-to-noise ratio be our discriminatory measure; this allows us to detect the part of the noise process affecting the observed data that once removed suggests a better identification of the latent structure in the volatility process.

The experiments in this section are conducted with the *wavelets module* written in *S-Plus* by Bruce and Gao (1994). Our goal is thus to investigate whether de-noising the data with wavelets and applying the waveshrink estimator can improve the ability to detect the latent volatility structure characterizing the observed time series. We compare model performances for original and transformed time series and use the best model selected among many others tested. The original series

is indicated by SER1, and the other series derived from the application of the waveshrink algorithm with non-decimated wavelets to the original series, first, and a normalization procedure after; we indicate the second series by SER4.

Many volatility models have been tested and the one with the best fit was found to be a GARCH(1,1), while the conditional variance for a GARCH(p,q) process is given by:

$$h_t = a + \sum_{i=1}^{p} a_i f_{t-i}^2 + \sum_{i=1}^{q} b_i h_{t-i}.$$ (9)

The conditional mean equation, which in Eq. (6) was indicated as a semi-parametric regression model, can be further modelled in its term f_i; in particular, we allow for a lag or an equivalent MA(1) term, to account for the influence of lagged residuals, and for a regressor matrix that includes exogenous variables designed to limit the inevitable residual misspecification, i.e. holiday and weekend dummies, the series of index levels, running means computed over 5, 7 and 15 days and the correspondent running volatilities over 5 and 7 days (all of these variables are selected among others on the basis of significance tests). The model thus should be written with a conditional mean equation like this:

$$y_i = \text{const.} + Z_i + \epsilon_i$$ (10)

where, in $Z_i = \alpha X_i + \beta f_i$, the **X** matrix includes the exogenous variables described before and the signal vector includes the error f_{i-1}, i.e. the MA(1) term.

Since the Gaussian GARCH models do not seem to completely capture the degree of leptokurtosis observed in the data, by leaving residuals with a clear evidence of a heavy tailed conditional distribution, a Student's t $g(x) = c \frac{1}{(1+\frac{x^2}{v-2})^{\frac{v+1}{2}}}$, where $c = \frac{\Gamma(\frac{v+1}{2})}{(\pi(v-2))^{\frac{1}{2}}\Gamma(\frac{v}{2})}$ is selected; note that $x = f_i h_i^{-\frac{1}{2}}$, i.e. the standardized residuals ξ_i, and that the degrees of freedom are estimated with the other parameters, say Θ, in the model.

A leverage term is also inserted in the conditional variance equation and is estimated to allow for the consideration of asymmetric effects of positive and negative returns. This term is therefore to be accounted, together with the GARCH structure, for measuring the combined effects. With the leverage term included, a conditional variance equation assumes the following form:

$$h_t = a + \sum_{i=1}^{p} a_i(|f_{t-i}| + \gamma_i f_{t-i})^2 + \sum_{i=1}^{q} b_i h_{t-i}.$$ (11)

7.1. *Estimation and prediction performance*

The selected model was estimated over the two series available by using the *S-Plus GARCH module* [10]. We used the entire set of data for estimating the model. By the prediction error decomposition, the log-likelihood function for a sample

Table 4. Estimated parameter values for the time series, along with the t statistics (in parenthesis).

Parameters	SER1	SER4
MA(1)	−0.017	0.618
	(−1.97)	(108.48)
B4	−0.945	0.243
	(−11.17)	(19.54)
B5	1.994	0.493
	(30.58)	(46.18)
B6	−0.086	0.099
	(−1.82)	(17.04)
lev	0.118	0.201
	(7.64)	(4.65)
ARCH	0.053	0.691
	(7.13)	(22.95)
GARCH-1	0.763	0.212
	(7.80)	(11.81)
GARCH-2	0.061	0.036
	(0.72)	(4.33)

$y_1, \ldots, y_i$ is given by $l_i(\Theta) = \log L_T(\Theta) = \sum_{i=1}^{T} \log f(y_i|\Psi_{i-1}) = -\frac{1}{2} \sum_{1=1}^{T} \log h_i + \sum_{i=1}^{T} \log g(\frac{f_i}{h_i^{1/2}})$, where $g(\cdot)$ is the Student's t distribution described before. The maximum likelihood estimation procedure gives the following values for the four series investigated: −14485 (SER1) and 9980 (SER4).

Our estimation analysis is summarized in Table 4. In short, these results indicate that the *MA(1)* coefficient, inserted into the conditional mean equation to take into account the one-step behind residuals, increases its absolute value by going from SER1 to SER4, and becomes significant. The *leverage* term, i.e. *lev*, is relatively small in its absolute value and only modestly significant. Note that the running means (B4 for 7 days, B5 for 5 days and B6 for 15 days) and volatilities inserted in the conditional mean equation are computed by $run.mean_{n,t} = \frac{1}{n} \sum_{t-n+1}^{t} r_t$ and $run.vol_{n,t} = \sqrt{\frac{1}{n-1} \sum_{t-n+1}^{t} (r_t - run.mean_{n,t})^2}$, where r_t are the returns.[e]

The significant *MA(1)* might indicate misspecification, bringing to the conclusion that profit-taking strategies are easily available to market agents in the long run. It is reasonable that these profit opportunities wouldn't stay undetected for long periods of time, thus explaining the fact that a not significant value would be more in line with the market efficiency theory. However, recent empirical findings in market microstructure studies address many possible sources of correlation in

---

[e]We report only some of these values in the table, having the excluded variables resulted not significant across experiments.

observed returns, and the impact of several factors should be taken into account. Therefore, it might be that instead of just thinking that de-noising can destroy the characteristics of the observed data and thus emphasize features unknown before, but contradicting the proper market behaviour, de-noising might indirectly confirm the presence of non-efficiency in the market. This condition might not be limited to the short term, due to the possible presence of regimes which remain hidden when analysing the original data or the presence of long run dependence, together with the short run easily uncovered.

We also note that pure ARCH effects show up more clearly when the decimated and ST waveshrunken estimators are applied to the data; their absolute value increases and they become even more significant. The GARCH coefficients behave in a different way: the GARCH-2 component, which we included for being able to capture some more dependency, is small and not significant for SER1, while is only modestly significant for SER4. For this reason we do not consider its relevance negligible and include it in our model specification; the GARCH-1 coefficient reduces its absolute value when computed for SER4, remaining sufficiently significant.

At first sight, we observe that a more precise isolation of the pure ARCH effects in the structure of volatility can be important in those circumstances when their presence could be easily questioned, due to particular complex dynamics characterizing the underlying stochastic processes. Moreover, the fact that past volatilities become less important with waveshrunken series in determining the current value of the same variable can suggest that an effective separation of noise and signal helps for understanding how the recursive effects propagate both temporally and spatially.[f] The investigator ignores, of course, *a priori* what is the best, as far as the influence of these delayed effects on the most recent observation is concerned, but it makes sense to consider the fact that even though with autoregressive dynamics one can hope to predict better, in high volatility market phases the abovementioned propagation effects would probably prevent the analyst from effectively detecting the true latent features, due to the dominant role that noise would have under these circumstances.

Since predicting the structure of volatility is our goal here, we also report plots of the predictions obtained for the series of volatility values. We have some evidence that once the noise in the observations is reduced, the forecasted volatility function, despite the evidence of some diminutive power compared to the squared returns series, is such that the gap between the two series is sensibly reduced and they appear more similar in the pattern they follow. Therefore, the de-noised squared realized returns can be a better indicator to track the variability of the latent volatility function. We compute the out-of-sample predictions as follows: we first fit our model to 13490 observations, for every series, and predict one step ahead with the selected GARCH model, i.e. we make a forecast for the observation at time

---

[f]In this last case via the so-called volatility clusters, which can be detected by the observation of occasional but temporally persistent bursts of activity in the data.

$t = 13491$; next we update the sample by using 13491 observed values, re-fit the model, predict the value at $t = 13492$, and so forth until we predict at $t = 13505$. We end up with a small sample of 15 one step ahead predictions with which we compare the prediction power of our estimated model on the original and waveshrunken series. The correspondent plots of realized squared returns vs. volatility forecasts are reported, together with some diagnostic plots and measures.

One can observe in Fig. 7 that the waveshrunken estimators do a good job in emphasizing the presence of GARCH effects in the squared de-noised returns. While the autocorrelation functions of the squared standardized residuals still report evidence of structure left, due to a certain degree of misspecification in the conditional mean equation, the Ljung-Box test statistic Q computed for the squared standardized residuals turns out to be 10.49 and 5.53 for respectively SER1 and SER4. Therefore the two series reveal that we cannot reject at either 1% or 5% levels the null hypothesis of white noise residuals, being $Q \sim \chi^2_{12}|H_0$.

We also compared the one-step ahead forecasted volatility with the squared realized returns (see Fig. 8), and notice a higher variability of the prediction curve for the waveshrunken series, which follows more closely the pattern of the squared returns dynamics. This conversely means that squared returns can investigated more usefully to understand the latent volatility behaviour. In terms of quantitatively measuring the prediction performance, we compute the *Root Mean Square Error*, i.e. $RMSE = (\frac{1}{T} \sum_{i=1}^{T} [y_i - \hat{y}_i]^2)^{\frac{1}{2}} = (\frac{1}{T} \sum_{i=1}^{T} e_i^2)^{\frac{1}{2}}$, which gives the values 2.26686 for SER1 and 0.60964 for SER4[g].

It thus seems that some latent structure previously hidden has been detected; one reason for this is that by limiting the influence of our measurement error we manage to better separate noise and signal structures, hence our squared returns can be better estimators for the underlying latent volatility. Literature reporting experiments done with higher frequency data [1] show how increasing the frequency of observation reduces noise and improves volatility prediction. We obtained similar results for daily data, directly de-noising them.

Returning to the correlation appearing from the estimates of Table 4, the non-vanishing *MA(1)* term seems to only apparently violate the expected picture of a rapidly decaying autocorrelation function of returns, since by looking at the same function computed for the squared de-noised series (Fig. 7) one may note that there is a slow decay instead, thus emphasizing the presence of possible long memory, which thus should be further investigated before judging definitely the meaning of the significant moving average coefficient. The autocorrelation function of the squared standardized estimated residuals leaves the reader with the same impression, i.e. that something of the structure of both the series is not completely captured, and only with the de-noised return series is the same idea is clearly conveyed [Fig. 7(c)].

---

[g]In the RMSE formula, T is the number of predictions and the y and $\hat{y}$ are respectively the squared realized returns and the GARCH predicted conditional variances.
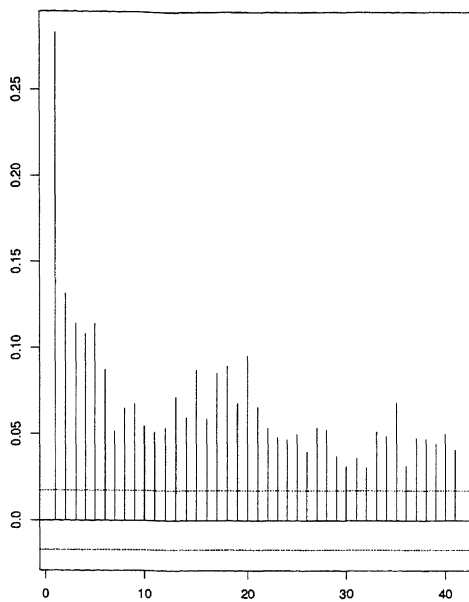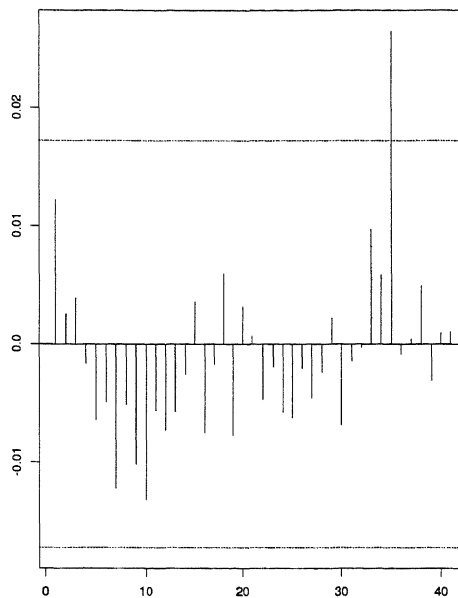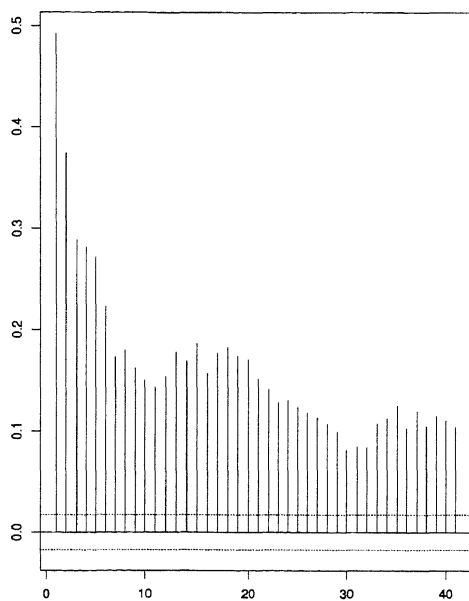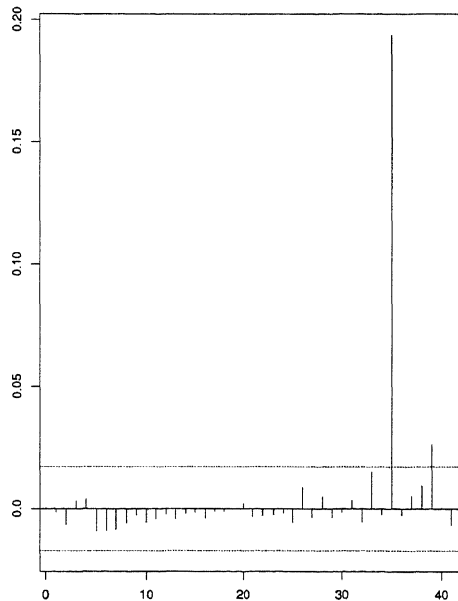
Fig. 7. (a) ACF squared SER1; (b) ACF squared st. residuals from GARCH estimates on SER1; (c) and (d) show in the same sequence results for SER4.
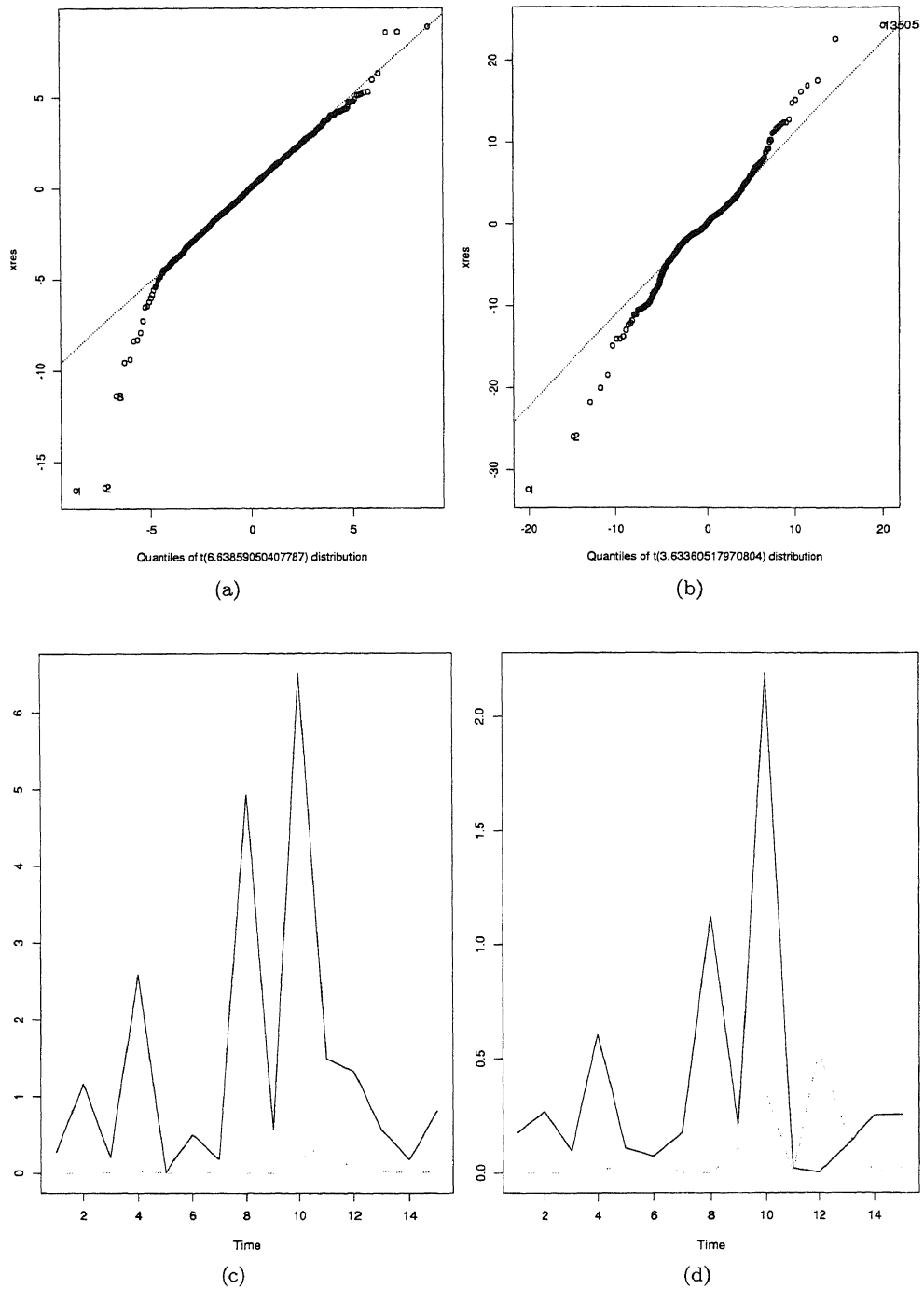
Fig. 8. (a) and (b) QQ plot st. residuals from GARCH applied respectively on SER1 and SER4; (c) and (d) comparison of squared returns and volatility predictions with GARCH applied on respectively SER1 and SER4.

8. Conclusions and Future Directions

We studied the impact of Wavelet Multi-resolution analysis on financial time series and showed that it offers interesting insights for discovering the presence of volatility structure at various resolution levels. In financial time series it's possible that information affects the market at different time horizons, such that its effects on returns can correspondingly be analyzed at different time scales and frequencies. We observed that (1) signal structure shows up mostly at very fine scales (2) the number of wavelet coefficients is big and thus the risk of overfitting is high when modelling directly with them (3) wavelet packets suggest rich dictionaries of functions from which a good basis can be selected, via best basis or matching pursuit algorithms (4) denoising the series via the wavelet shrinkage algorithm allows for a substantial reduction in the number of coefficients, thus suggesting a more selective signal reconstruction. We show results about modelling with GARCH when the data are noisy and when they are pre-processed via wavelet transforms. A better volatility prediction power for one step ahead forecasts arises in the case of de-noised data, thus indicating that latent volatility features can be better detected. This result is usually achieved when less measurement noise is allowed through an higher sampling frequency for the observed signal.

Acknowledgment

References

[1] T.Andersen and T.Bollerslev, *Answering the Critics: Yes, ARCH Models do Provide Good Volatility Forecasts*, NBER Working Paper (1997), p. 6023.

[2] T.Bollerslev, *Generalised autoregressive conditional heteroskedasticity*, J. Econometrics **31** (1986) 307–327.

[3] A.Bruce and H.V.Gao, *S+Wavelets*, StaSci Division, MathSoft Inc., Seattle (1994).

[4] E. Capobianco, *Wavelets for High Frequency Financial Time Series*, Interface 1999 Proceedings, Schaumburg (2000), pp. 373–378.

[5] R. Coifman and V. Wickerhauser, *Entropy based algorithms for best basis selection*, IEEE Transactions Information Theory **38** (1992) 713–718.

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

[7] D. Donoho and I. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. American Statistical Association (1990) 1200–1224.

[8] R. F. Engle, *Autoregressive Conditional heteroskedasticity with estimates of the variance of the UK inflation*, Econometrica **50** (1982) 987–1008.

[9]   S. Mallat and Z. Zhang, *Matching pursuit with time frequency dictionaries*, IEEE Transactions Signal Processing **41** (1993) 3397–3415.

[10] D. Martin, H. V. Gao and Z. Ding, *S+GARCH, Data Analysis Product Division*, MathSoft Inc., Seattle (1996).

[11] I. Meyer, *Wavelets: Algorithms and applications*, SIAM, Philadelphia (1993).

[12] U. A. Muller, M. M. Dacorogna, R. D. Dave', R. B. Olsen, O. V. Pictet and J. E. von Weizsacker, *Volatilities of different time resolutions — Analyzing the dynamics of market components*, J. Empirical Finance **4** (1997) 213–239.

[13] D. B. Nelson, *Filtering and forecasting with misspecified ARCH models I*, J. Econometrics **52** (1992) 61–90.

[14] A. S. Weigend and N. A. Gershenfeld, *Time series prediction: Forecasting the future and understanding the past*, Proceedings Vol. XV, Santa Fe Institute, Studies in the Science of Complexity, Addison-Wesley (1994).