Eur. Phys. J. B **27**, 201–211 (2002)
DOI: 10.1140/epjb/e20020154

THE EUROPEAN PHYSICAL JOURNAL B
EDP Sciences

Hammerstein system represention of financial volatility processes

E. Capobianco[a]

CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Received 31 December 2001

Abstract. We show new modeling aspects of stock return volatility processes, by first representing them through Hammerstein Systems, and by then approximating the observed and transformed dynamics with wavelet-based atomic dictionaries. We thus propose an hybrid statistical methodology for volatility approximation and non-parametric estimation, and aim to use the information embedded in a bank of volatility sources obtained by decomposing the observed signal with multiresolution techniques. Scale dependent information refers both to market activity inherent to different temporally aggregated trading horizons, and to a variable degree of sparsity in representing the signal. A decomposition of the expansion coefficients in least dependent coordinates is then implemented through Independent Component Analysis. Based on the described steps, the features of volatility can be more effectively detected through global and greedy algorithms.

PACS. 02.50.Tt Inference methods – 05.45.Tp Time series analysis – 02.60.Gf Algorithms for functional approximation

1 Introduction

One goal of this paper is to show that financial time series analysis could take advantage from some methodological principles adopted in other research fields, where the most important goal is to detect, through sparsely represented signals, all the relevant information, in the form of factors or features, which characterize natural and experimental phenomena [14,30,35]. Due to the importance of separating the underlying structure of volatility from the noise, and thus to isolate the role that non-stationarity can play, adaptive estimators are needed, in particular when dealing with classes of functions which are endowed with a variable degree of smoothness. Therefore, for time non-homogeneous classes of functions like those related to the realizations of volatility processes, inference models are required to be flexible, in terms of the assumptions about the probability distributions involved, and non-parametric estimators are especially useful.

Another aspect of interest in this study, is that given a certain stochastic process, the decomposition of its observed realizations in statistically independent coordinates may be a key goal in applications. A possible way of proceeding is to look at the problem of reducing the dependence structure by developing techniques which transform the data. Wavelets can play an important role in these last cases, since they yield [1,27] de-correlating and stationarizing effects on the computed coefficient sequences. Consequently, statistical inference can be more effective in the projected domain, yielding near-optimal minimax estimates.

Our experiments show that approximation algorithms such as the *Matching Pursuit* (MP) [31] effectively detect patterns and features in high frequency financial time series [6], and can be successfully combined with an *Independent Component Analysis* (ICA) [8,12] with the aim of artificially learning the structure of a volatility process.

2 Hammerstein systems

The theory of identification of linear and non-linear dynamic systems offers many aspects interconnecting disciplines like signal processing, information theory and statistical inference. *Hammerstein* systems [20,23,29] are a possible model for dealing with both a non-linear dynamic sub-system (NLDS) and a linear (LDS) one, represented in time varying cascade time series form, with parametric or non-parametric statistical structure, and under stationary or non-stationary conditions. The model can be simplified as follows:

$$X_t \rightarrow NLDS \rightarrow Z_t \rightarrow LDS \rightarrow Y_t$$

where Z_t is not observable. The main problems are 1: recovering the unknown non-linearity from the observed time series, for which kernel estimators have been usually employed, and 2: searching for localization and

[a] e-mail: Enrico.Capobianco@cwi.nl

parsimony/sparsity in the functional representation. A compact functional form for the Hammerstein system is:

$$Y(t) = \int_{-\infty}^{\infty} \lambda(t-\tau)K(X(\tau))\mathrm{d}\tau \tag{1}$$

where K indicates the NLDS and λ the LDS blocks. A correspondent discrete time series representation for (1) is given by the following functional relation, where a time-varying, truncated and with finite time support structure may be described by a *signal + noise* model:

$$Y(t) = \sum_{i=1}^{N}\sum_{\gamma=0}^{l} \lambda_i(t,\gamma)K_i[X(t,\gamma)] + \epsilon(t) \tag{2}$$

where now the functions and the time varying kernels as well are not specified, and the ϵ term represents the disturbance. This system, depending on the specification of the λ and K dynamics, and on the ϵ's *pdf*, can be regarded as a parametric, semi-parametric or non-parametric statistical model, and it can be adapted to the analysis of volatility processes. Among the possible characterizations, we have chosen one that requires two fundamental steps: wavelet-based transforms and independent component decompositions, as explained below.

3 Independent component analysis

ICA is related to linear and nonlinear mixture models [2], and refers to noise-free or noisy data applications, thus reflecting also typical choices (mixture models) and empirical evidence (noise effects) found in financial modeling. Furthermore, one might find more convenient to work with innovation processes [25] derived from conditional values of observation processes, because the latter are usually more independent and non-Gaussian. Volatility processes appear as innovation processes adapted to filtered information flows and are characterized by an high order dependence structure.

With Gaussian signals the *Independent Components* are the *Principal Components*, while with non-Gaussian signals ICA delivers superior performance, due to the fact that it relies on high order statistical information coming from higher order moments of the probability distribution. This last fact leads to applications in financial time series, due to the non-Gaussianity of return distributions, even if other reasons may deserve consideration.

A non-linear relation $Y = f(X)$ may be linearly approximated by AS, a linear ICA (L-ICA) where A is an $N \times M$ mixing matrix, with $N \geq M$[1], and where the sensors $X_i, i = 1, \ldots, N$ combine sources $s_i, i = 1, \ldots, M$, which are independent, non-Gaussian, but also unknown. Thus, the problem is to estimate both the sources and the mixing matrix from the observed outputs.

One thus tries to find a linear transform or de-mixing matrix W such that $U = WX = WAS$ are delivered as the most independent or least dependent estimates for the unknown sources. If $W = A^{-1}$, clearly $U = S$, up to permutation (R) and scale (D), *i.e.* $P = RD = WA$, and thus with $P = I$ one gets perfect separation under normalization and re-ordering.

The solutions, in real applications, hold only approximately, and are given by algorithms such as the *Joint Approximate Diagonalization of Eigenmatrices for Real signals* (JadeR) [9] or the *fastICA* [24,26], just to quote the most popular ones. The de-correlation and rotation steps which have to be implemented, thus deliver a set of approximate M independent components. The following examples combine model structures of the described systems;

Hammerstein system and L-ICA:

The system $Y_t = V_t + \epsilon_t$, with $V_t = \sum_m \lambda_m Z_{t-m}$, represents an Hammerstein system output relation, but also an example of noisy ICA. If $Z_t = K(X_t)$ works through some wavelet-based transforms, then the L-ICA acts on the sequence space formed by the wavelet expansion coefficients and decomposes the scales or resolution levels.

Latent variable system:

The following extension, $Y_t = A_t Z_t + \epsilon_t$; $Z_t = C_t \Phi_t + \eta_t$, where, as before, $Z_t = K(X_t)$ and Φ_t is an approximation dictionary, refers to a sparsity constraint embedded in the system, together with a decomposition step for Z_t, due to the presence of an atomic set of approximating functions collected in an overcomplete dictionary. The last representation suggests that the time varying Hammerstein series can be approximated through the specification of its kernels as a linear combination of basis sequences ϕ_k, endowed with certain time and frequency supports, depending on the selected atoms. The basis approximation of the Hammerstein kernel can be described as:

$$\lambda_i(t,\gamma) \approx \sum_{s=0}^{S} \alpha_i(s,\gamma)\phi_s(t) \tag{3}$$

which leads to the following representation, in replacement for (2):

$$Y(t) = \sum_{i=1}^{N}\sum_{\gamma=0}^{l}\sum_{s=0}^{S} \alpha_i(s,\gamma)\phi_s(t)K_i[X(t,\gamma)] + \epsilon(t). \tag{4}$$

Identifying such system requires the estimation of the expansion coefficients α; this problem is faced in this paper, too. Under simplifying conditions on the functions involved and by re-arranging variables and parameters in (4), we can address with $Y = X\beta + E$ the compact

[1] An $M \times M$ mixing matrix A is often studied; usually $M \ll N$, with N the number of sensor signals. However, $M = N$ holds in many cases, too.

matrix form of (4), with E_t representing both the measurement and the approximation errors. This system can be solved by numerically stable routines which compute pseudo-inverse transform of X; from the estimated β coefficients one can thus compute the approximated time varying Hammerstein kernels.

Note that a function approximation problem is also addressed, since there are functional components in β that have to be approximated from a finite set of available training examples. Furthermore, the estimates for the parameters in the model are going to be computed under almost no assumptions, since both non-Gaussianity and non-stationarity are likely conditions under which the system dynamics operate. The same conditions are typically found while empirically investigated volatility processes.

4 Sparse component analysis

4.1 Modeling volatility processes

We start by casting the volatility process of interest in a very general frame [7], already introduced in the examples, and so to represent its dynamics by the following linear system of equations:

$$Y_t = A_t X_t + \epsilon_t \tag{5}$$
$$X_t = C_t \Phi_t + \eta_t \tag{6}$$

where Y_t are the observed financial returns[2], X_t are *unknown system sources*, A_t is an *unknown mixing matrix*, $\epsilon_t \sim i.i.d.\ (0, \sigma_{\epsilon,t})$ is a noise process. Note that $v_t = \sigma^2_{\epsilon,t}$ is here considered the volatility process, *i.e.* the latent process underlying the observed returns.

The sources X_t have a possibly *sparse* decomposition through Φ_t, a selected dictionary of functions delivering either a basis or an *overcomplete representation* [10,30] for the signal under investigation. In the latter case, linear combinations of elements may suggest possible representations of remaining dictionary structures, thus offering non-unique signal decompositions.

The corresponding expansion coefficients are here indicated by C_t, while η_t is an i.i.d process, with no constraints on the probability distributions[3]. Therefore, it may hold as a quite general frame and it suggests a sort of model-free approach for representing the dynamics of the system of interest.

A special case [35] is when a *dual system* can be formed, *i.e.* when a basis is obtained; in that case the system can change according to the transform $\Phi_t^{-1} = \Psi_t$. As a direct consequence, $X_t \Psi_t = C_t \Phi_t \Psi_t + \eta_t \Psi_t$; this last expression can be expressed equivalently as $\tilde{X}_t = C_t + \tilde{\eta}_t$, while at the observation level $Y_t = A_t C_t + A_t \tilde{\eta}_t + \epsilon_t$, or also $Y_t = A_t C_t + \xi_t$, with $\xi_t = A_t \tilde{\eta}_t + \epsilon_t \equiv A_t \eta_t \Psi_t + \epsilon_t$.

To summarize, a new system is found:

$$\tilde{X}_t = C_t + \tilde{\eta}_t \tag{7}$$
$$Y_t = A_t C_t + \xi_t. \tag{8}$$

If the *signal-to-noise ratio* (S/N) with regard to the sources is high, then $\eta_t \approx 0$ and $\xi_t = \epsilon_t$. Thus, the same volatility process initially described is found. If instead S/N is low, the (square root) volatility process becomes characterized by $\Sigma_t = D_t + \sigma_{\epsilon,t}$, where $D_t = A_t \sigma_{\eta,t} \Psi_t$. In the latter case, an overcomplete dictionary is available. Note that system (5-6) represents a volatility process in a way that generalizes other typical autoregressive form of dependence employed by many models, depending on the structure of the Φ_t matrix[4]; the volatility structure is expressed in a non-parametric form and is investigated by selected dictionaries of functions[5].

Conversely, in system (7–8) the original returns have a new decomposition with (8), where the mixing matrix A_t operates on the computed transform expansion coefficients C_t. While A_t accounts for modulating the dependence structure of the latent volatility sources, the wavelet-based expansion coefficients become the inputs for the ICA step that follows. These coefficients, in (7), are now sparsely represented in transformed and scaled sources of volatility information obtained through ICA.

In other words, one can work in a signal or a sequence space, respectively with functions or coefficient sequences; the choice may depend on criteria such as sparsity of representation and statistical independence of the coordinates under which the stochastic process is decomposed. As a rule, we can also build an optimization system with a regularized objective function through some smoothness priors, so to estimate the parameters involved. The strategy here is to proceed recursively (in the mean square sense), through iterations of a MP processing over the observed returns and with the WP libraries, thus looking at $Y_t \approx P_t \Phi_t + \xi_t = A_t C_t \Phi_t + \xi_t$[6]

The MP algorithm works on a sparse P_t and performs a denoising step, but remains unable to disentangle the components composing the operator P_t. It will be left to an ICA step dealing with this aspect; the compression and the decorrelation properties of wavelet transforms can be better supported with a more effective search for least dependent components.

[2] Stock returns are computed in the usual way, as $r_t = \ln(p_t/p_{t-1}) \times 100$, where p_t are the prices of shares, indexes, commodities or other financial activities.

[3] Thus the fact that we don't require positivity means that we are not describing volatility through equation (6), but simply sources of it.

[4] We might also design a state-space structure for representing the system dynamics.

[5] We can also maintain, according to the representation adopted, an underlying well-known hypothesis that a mixture basic law of information arrivals is governing the market dynamics.

[6] The noise is including an approximation error from the system equation and residual measurement effects ϵ_t.

4.2 Searching for sparsity

With *Sparse Component Analysis* (SCA) [14], one attempts to optimize the compression power of certain transforms and to attain estimation results under non-parametric statistical models, according to minimax results in decision science.

Based on system (5–6), alternative learning models can be designed; for instance, A and C can be computed from the following optimization problem:

$$\min_{A,C} \frac{1}{2\sigma^2} \mid\mid AC\Phi - X \mid\mid_P^2 + \sum_{j,k} \beta_j h(c_{j,k}) \qquad (9)$$

where a connection with sparse representations [19] can be made by changing the norm in P and allowing for more robustness in the underlying probability distributions. With $h(.)$ representing a prior distribution on the dictionary expansion coefficients, or otherwise an empirical probability distribution function that could be computed from the estimated wavelet coefficients, the functional (9) generalizes other similar structures like the *Method of Frames*, the *Basis Pursuit*, or the equivalent *Linear Programming* problem representations (see [10] for a review).

As a final remark of this section, note the term $AC\Phi$ can be replaced, for $M = N$ and the de-mixing matrix indicated by $B = A^{-1}$. Thus, it follows that $S \approx BX$ and the term within the norm of the objective function becomes $\mid\mid C\Phi - BX \mid\mid_P^2$.

The goal of achieving sparsity by using wavelet-based representations of signals inspires strategies to eliminate the redundant information. This can be done in the wavelet coefficients domain, given the relation between true and empirical coefficients, $\tilde{d}_{jk} = d_{jk} + \epsilon_t$. The *wavelet shrinkage principle* [15–17] applies a thresholding strategy which yields de-noising of the observed data; it operates by shrinking wavelets coefficients toward zero so that a limited number of them will be considered for reconstructing the signal.

Since a better reconstruction might be crucial for financial time series in order to capture the underlying volatility structure and the hidden dependence, de-noising could be usefully employed for these temporally and spatially heterogeneous signals.

5 Multiresolution analysis and overcompleteness

In order to build a wavelet system one needs (**A**) a *scaling function* ϕ whose dilates and translates constitute orthonormal bases for all those V_j subspaces which are obtained as scaled versions of the subspace V_0 to which ϕ belongs, and (**B**) a *mother wavelet* ψ together with ψ_{jk} generated by j-dilations and k-translations, such that $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$. Furthermore, with all of the information obtained from the approximations computed at successively coarser resolution levels we can form [13] a *Multiresolution Analysis* (MRA), *i.e. a sequence of closed subspaces*[7] *satisfying* $\ldots, V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \ldots$, *with* $\bar{\cup}_{j\in Z} V_j = L_2(R)$, $\cap_{j\in Z} V_j = \{0\}$ *and the additional condition* $f \in V_j \Longleftrightarrow f(2^j.) \in V_0$.

The last condition is a necessary requirement for identifying the MRA, meaning that all the spaces are scaled versions of a central space, V_0. An MRA approximates $L_2[0,1]$ through V_j generated by orthonormal scaling functions ϕ_{jk}, where $k = 0, \ldots, 2^j - 1$. These functions allow also for the sequence of 2^j wavelets ψ_{jk}, $k = 0 \ldots, 2^j - 1$ to represent an orthonormal basis of $L_2[0,1]$. The set of shifted scaling functions $\{\phi_0(t-k), k \in Z\}$ is an unconditional Riesz basis for V_0, *i.e.* linearly independent functions are obtained; then, the scaled and shifted functions $\phi_{jk}(t)$ are Riesz bases as well for the scaling spaces V_j. On these spaces the incremental information process is due to the signal which is projected such that $P_{V_j}X(t) = \sum_k c_x(j,k)\phi_{j,k}(t)$ and $D_j(t) = P_{V_{j-1}}X(t) - P_{V_j}X(t)$, or otherwise directly through the projections in the W_j wavelet subspaces, *i.e.* $D_j(t) = P_{W_j}X(t) = \sum_k d_x(j,k)\psi_{j,k}(t)$.

Signal decompositions with the MRA property have near-optimal properties in a quite wide range of inhomogeneous function spaces [13,22,32].

As an alternative to orthonormal wavelet representations there are other signal representations defined as overcomplete, and able to offer some advantages (see [34], for details), particularly in terms of robustness to noise in the coefficients and to quantization effects, and also with regard to the aspects related to a certain freedom in choosing the wavelet family and exploiting the irregular sampling design. A practical example comes from function dictionaries, *i.e.* collections of parameterized atomic structures [10]; they are available for representing many classes of functions and are formed directly from a particular family, like wavelets, or from merging two or more dictionary classes.

Dictionaries which are overcomplete bring of course redundancy into the model and deliver non-unique signal decompositions. When instead a basis can be selected, the dictionary results complete. The kind of overcomplete representation that we have adopted in the experiments is based on *wavelet packets* (WP), which represent an extension of the wavelet transform to a richer class of building block functions. They allow for a better adaptation due to an oscillation index f related to a periodic behavior in the series which delivers a richer combination of functions. As for wavelets, an admissibility condition is required, too: $\int_{-\infty}^{+\infty} W_0(t)\mathrm{d}t = 1$, $\forall (j,k) \in Z^2$ we have from [28]:

$$2^{-\frac{1}{2}} W_{2f}(\frac{t}{2} - k) = \sum_{i=-\infty}^{\infty} h_{i-2k} W_f(t-i) \qquad (10)$$

[7] Here expressed in nesting order as in a ladder of Sobolev spaces, with the more negative the index the larger the space.

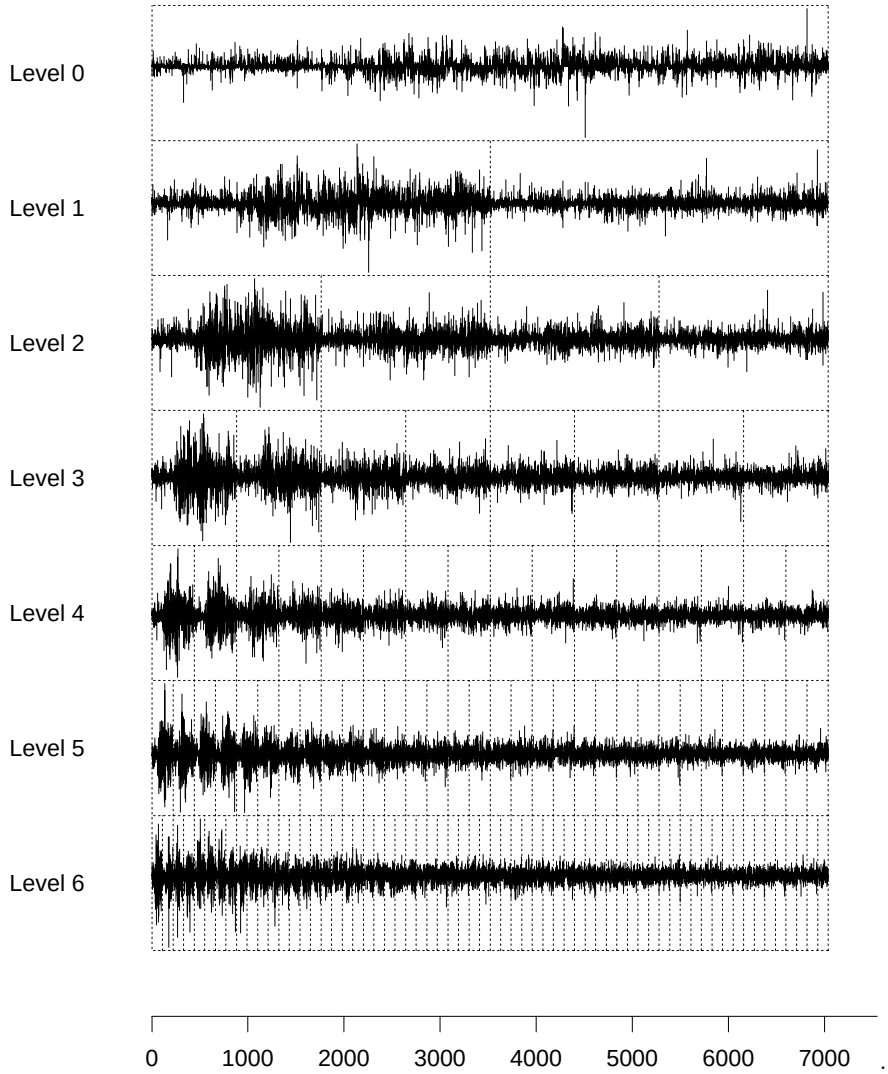

Fig. 1. WP table and time-frequency signal segmentation obtained resolution-wise or level-by-level.

where f relates to the frequency and h to the low-pass impulse response of a quadrature mirror filter, and

$$2^{-\frac{1}{2}} W_{2f+1}\left(\frac{t}{2}-k\right) = \sum_{n=-\infty}^{\infty} g_{n-2k} W_f(t-n) \qquad (11)$$

where g is an high pass impulse response. For compactly supported wave-like functions $W_f(t)$, finite impulse response filters of a certain length L can be used, and by P-partitioning in (j,f)-dependent intervals $I_{j,f}$ one finds an orthonormal basis of $L^2(R)$ (*i.e.* a wavelet packet) through $\{2^{-\frac{j}{2}} W_f(2^{-j}t-k), k \in Z, (j,f) \mid I_{j,f} \in P\}$. A better domain, compared to simple wavelets, is obtained for selecting a basis to represent the signal, but an orthogonal wavelet transform can always be selected by changing the partition P and defining $w_0 = \phi(t)$ and $W_f = \psi$.

Figure 1 describes the structure of within-block coefficients of the WP formulation, and thus represents its contribution in representing the signal features under a varying oscillation index. The WP table presents coefficients stored in sequence order according to increasing oscillation index; the blocks are ordered by frequency, and within the blocks the wavelet coefficients are ordered by time. Thus, the low frequency information in the signal is expected to be concentrated at the left and the high frequency information at the right of the table[8].

5.1 The matching pursuit learning algorithm

The MP algorithm is an example of a greedy approximation algorithm that has been successfully implemented in many studies for its simple structure and effectiveness. A signal is decomposed as a sum of atomic waveforms, taken from families such as Gabor functions, Gaussians, wavelets, wavelet and cosine packets, among others. We focus on the WP table, and represent the signal as:

$$WP(t) = \textstyle\sum_{jfk} w_{j,f,k} W_{j,f,k}(t) + res_n(t).$$

[8] The oscillation index goes from 0 to $2^J - 1$, going rightwise.

This choice offers some advantages, like flexibility of the approximating kernels, better spatial adaptivity and time-frequency localization power, use of prior knowledge and possible dimension reduction.

In summary, the MP algorithm approximates a function with a sum of n elements, called atoms or atomic waveforms, which are indicated with H_{γ_i} and belong to a dictionary Γ of functions whose form should ideally adapt to the characteristics of the signal at hand. The MP decomposition refers to a greedy algorithm which 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 one allowing for the best fit. At each time step the following decomposition is computed, yielding the coefficients h_i which represent the projections, and the residual component, which will be then re-examined and iteratively re-decomposed according to:

$$f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + res_n(t) \tag{12}$$

1. initialize with $res_0(t) = f(t)$, at $i = 1$;
2. compute at each atom H_γ the projection $\mu_{\gamma,i} = \int res_{i-1}(t) H_\gamma(t) \mathrm{d}t$;
3. find in the dictionary the index with the maximum projection,

   $$\gamma_i = \mathrm{argmin}_{\gamma \in \Gamma} \mid\mid res_{i-1}(t) - \mu_{\gamma,i} H_\gamma(t) \mid\mid,$$

   which equals from the energy conservation equation $\arg\max_{\gamma\in\Gamma} \mid \mu_{\gamma,i} \mid$;
4. with the nth MP coefficient h_n (or $\mu_{\gamma_n,n}$) and atom H_{γ_n} the computation of the updated nth residual is given by:

   $$res_n(t) = res_{n-1}(t) - h_n H_{\gamma_n}(t);$$

5. repeat the procedure from step 2, with $n = n + 1$ and until $i \leq n$.

With $\mathcal{H}$ as an Hilbert Space, a function $f \in \mathcal{H}$ is decomposed in this frame as $f = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + Rf$, with f approximated in the g_{γ_0} direction, orthogonal to Rf, such that $\|f\|^2 = \mid \langle f, g_{\gamma_0} \rangle \mid^2 + \|Rf\|^2$. Thus, the minimization of the $\|Rf\|$ term requires a choice of g_{γ_0} in the dictionary such that the inner product term is maximized (up to a certain optimality factor). The selection of these atoms from the D dictionary is made by an index γ_0 based on a choice function conditioned on a set of indexes $\Gamma_0 \in \Gamma$.

5.2 The best orthogonal basis algorithm

The *Best Orthogonal Basis* (BOB) algorithm [11] is employed here as an alternative to the MP optimization method, with the goal of minimizing an additive[9] cost function computed within a library of orthonormal basis representations generated by the WP transform and through the correspondent expansion coefficients w_{jf}.

The procedure adaptively picks the best orthogonal basis among those which can be formed as sub-collections of the WP dictionary. The BOB algorithm thus represents a global optimizer which computes the transform by searching for the minimum of a cost function $E(C) = \sum_{j,f} E(w_{j,f})$ in O(LN) operations, with $L = \log_2 N$ the number of levels of the binary tree and N is the signal length (this compared to the O(MLN) cost of the MP, with M packets selected). The algorithm is known to deliver near-optimal sparsity representations, but it doesn't show the same property under harder conditions, like in non-orthogonal contexts.

In particular, the BOB algorithm steps find a minimum entropy transform from the dictionary, *i.e.* $\min\,[entr\ f(B)] \mid B \in \Gamma$, where B is an orthobasis in the selected dictionary Γ and $f(B)$ are a vector of coefficients in the same basis. In terms of the entropy, commonly used in statistics for estimation and compression problems, the cost function holds as $E^{ent}_{j,f} = \sum_k \hat{w}^2_{j,f,k}\ \log\ \hat{w}^2_{j,f,k}$, for $\hat{w}_{j,f,k} = w_{j,f,k} \times (\mid\mid\ w_{0,0}\ \mid\mid_2)^{-1}$. The total energy is given by $E = \sum_{k=1}^{n} f_n^2$, which in turn corresponds to decomposing the energy among details and approximations, *i.e.* $E^s_j + \sum_{j=1}^{J} E^d_j$, where $E^s_j = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^J}} s^2_{j,k}$ and $E^d_j = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^J}} d^2_{j,k}$, for $j = 1, \ldots, J$.

In Figure 2 we report in (A) the top-100 largest coefficients approximation with the BOB and the MP algorithms after running on the WP dictionary and show in (B) a comparison with the MP algorithm.

The locations of the high energy spots indicate different costs in terms of the computed entropy for the two dictionaries, depending on which frequency information is captured by the related transforms. A low frequency concentration of energy appears in the WP cost table. The plots suggest that BOB doesn't work optimally for the non-stationary signal, while MP works more efficiently; this is due to its greedy nature, and it results more effective for a better ability to capture the local features, both in time and in frequency. The MP scheme exploits the correlation power inherent to the collection of waveforms available through the WP dictionary, and it does so throughout more scales and by extending the basis which represents the signal.

6 Non-parametric statistical inference

The (covariance) non-stationarity of financial time series and their dependence structure are related aspects, in the sense explained by [4,33], especially when dealing with high frequency data.

We refer in our experiments to the Nikkei stock return index and choose the series of 1990, among several years of available market activity, with observations collected at high frequencies, *i.e.* every minute (1 min). The total sample has 35,463 data, with intra- daily trading prices covering the working week, holidays and weekends excluded.

[9] Non-additive cost functions and near-best bases can be considered too.

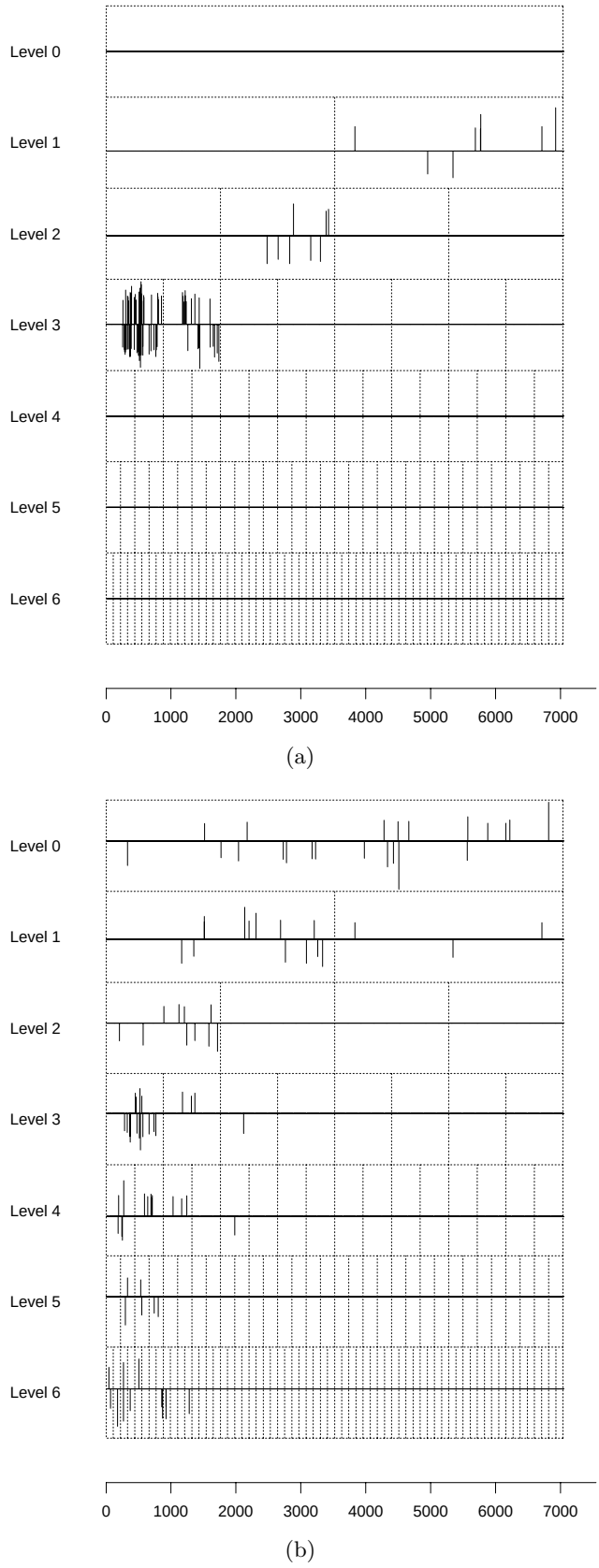

Fig. 2. Signal approximation from the WP table with the 100 largest selected coefficients from BOB (a) and from MP (b).

We then form a temporally aggregated time series of correspondent five-minute (5 min) data from the original one; thus, they are simple averages of components sampled at 1 min time interval. The aggregated sample consists of 7092 observations[10].

In Figure 3 we show some diagnostic plots, *i.e.* the autocorrelation functions computed on the 1 min and 5 min samples and considering absolute and squared values, from which the dependence structure clearly shows up. Volatility persistence is thus observed from these plots, and it is very likely that the dependence structures might be mixed with other features in the data, such as periodicity. These last components are usually not easily interpreted and detected by standard volatility models; thus, they may prevent the researcher from evaluating the underlying low frequency dynamics in the most correct way. Nevertheless, they might just represent, as also suggested by [33] the evidence of spurious features in the data.

6.1 Estimation procedure

Following [18], an oracle orthonormal basis $\mathcal{B}$ diagonalizes the covariance operator of a non-stationary stochastic process. Thus, one should estimate the covariance function Γ of the process by first rotating it into a basis $\mathcal{B}$, then forming the empirical covariance function $\hat{\Gamma}$ and finally eliminating the off-diagonal terms (following the expected values based on theoretical operators), in order to obtain after this last step $D_{\mathcal{B}} = \mathrm{diag}(\sigma^2_{i,\mathcal{B}})$. Then, by rotating back in the original basis, the resulting estimate is $\Gamma_{\mathcal{B}} = \mathcal{B} D_{\mathcal{B}} \mathcal{B}'$.

In order to start this procedure, a table of empirical variances σ_i^2 has to be built from the coefficients estimates obtained with *ad hoc* operators for these types of processes, *i.e.* cosine packets or localized cosines. The transformed values are then smoothed by thresholding, and the inverse transform is taken to achieve a reconstructed smoothed estimate $\bar{\sigma}_i^2$. At this point, the BOB algorithm is applied to a cost function built from the squared values of the smoothed empirical variances, and it yields the basis to be used for forming the diagonal matrix $D_{\mathcal{B}}$ and the covariance $\Gamma_{\mathcal{B}}$ estimates.

The problem with this elegant method is that only estimates for the conditional variance or the volatility may be used, instead of true values, which brings bias into the procedure. Furthermore, in our context, the covariances are time varying latent variables, either conditional on past return information sets or dependent on stochastic disturbances themselves. Therefore, special care should be required in order to deal with the selection of the thresholding rule. Our procedure considers these aspects, together with the fact that the setting is one with non-Gaussianity and non-stationarity.

The autocorrelation function instead of the autocovariance function is thus monitored as the key diagnostic function in our procedure, while it is computed based

[10] The experiments were conducted with $S+$ *Wavelets* [5].

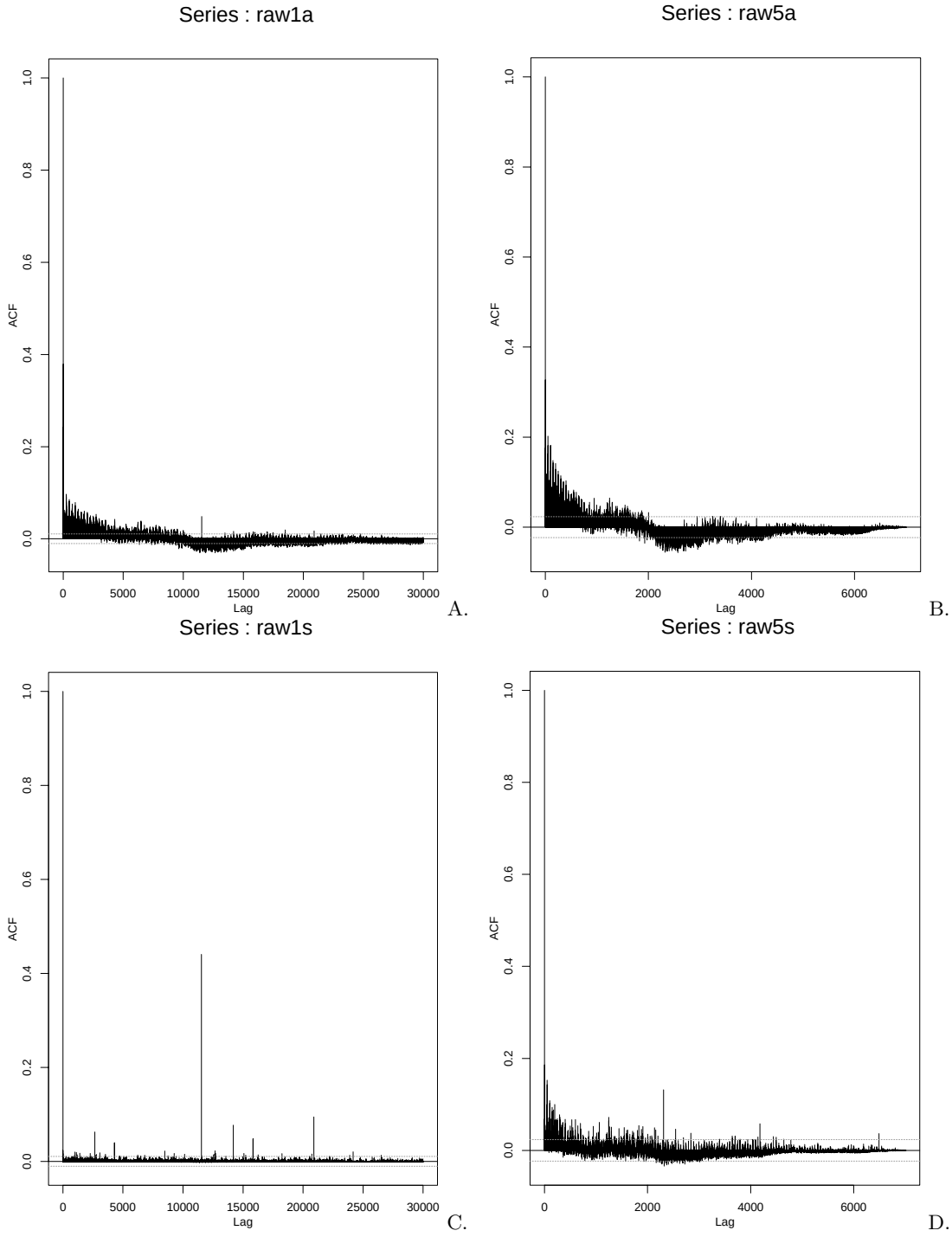

Fig. 3. ACF of absolute (indexed by a) and squared (indexed by s) raw 1 min and 5 min returns.

on the approximations delivered by the MP residues, through their absolute and squared values. De-noising, a by-product of MP, is enforced through the ICA whitening (decorrelation) step, and since a rotation (change of basis) is also performed in the expansion packet coefficients space, the dimension reduction occurs through the selection of the most important scales, according to the energy content of the diagonal entries of the estimated mixing matrix. The algorithm is described as follows:

- observe the features of $\mathcal{K}$, the original empirical ACF of the absolute and squared returns;
- compute $\bar{\mathcal{K}}$, by calculating its value based on the transformed residues of the MP approximations[11] obtained with an overcomplete dictionary;

[11] We have tested the MP approximation power by letting the algorithm work with 50, 100, 200 and 500 atoms from the selected WP dictionary.

Table 1. Weights of the estimated ICA mixing matrix distributed across resolution levels for residual 5 min series obtained in the WP table.

Res. lev.	0	1	2	3	4	5	6
WP-A							
level 0	**0.2218**	0.0028	0.0085	0.0047	0.0023	0.0069	0.0085
level 1	0.0002	**0.1951**	−0.0013	0.0001	−0.0189	−0.0035	−0.0037
level 2	0.0068	0.0003	**−0.167**	0.0015	0.0007	0.0019	−0.001
level 3	0.0031	−0.0057	−0.0008	**−0.1438**	−0.0019	−0.0045	0.0059
level 4	0.0012	−0.0125	0.0017	0.0028	**−0.1318**	0.0117	0.0
level 5	0.0032	−0.0023	0.0014	−0.0045	0.0008	−0.0011	**−0.1147**
level 6	0.0023	−0.0009	−0.0018	0.0047	−0.0082	**−0.121**	0.0017

– check how the features have been approximated by MP;
– select with ICA the most informative (in terms of energy) resolution levels from the computed MRA signals, thus reducing the dimension of the problem;
– re-start MP based on the new restricted range of scales and re-compute the transformed residues;
– get the final estimate $\hat{\mathcal{K}}$ and control the feature detection power

In Table 1 we have the estimated mixing matrices A, where the observed sensor signals are those computed at each resolution levels by the WP transform. These already de-seasonalized signals are now passed through the ICA algorithm for the extraction of "M" possible sources which we set equal to the number of sensors. For a possible interpretation of how these level dependent ICs may relate to financial market dynamics, activities and operations, one might consider that relevant work has been recently proposed by researchers addressing the hypothesis that financial markets operate under conditions driven by dynamics which are different according to the time horizons considered for evaluating returns.

Since our sensor signals are obtained from a multiresolution decomposition of the signal, instead of measuring each IC's contribution to the individual returns we extract from each detail level an approximate value suggesting its contribution to the signal features independently from the other levels. The highest values computed suggest what are the dominant ICs on a scale-dependent basis, without identifying their specific nature or the underlying economic factors, being them system dynamics or pure shocks.

From the WP estimated mixing matrix A we note a strong within-level factor always dominating apart from levels 5 and 6, where a mutual cross-influence appears to dominate. Considering the results obtained with the ICA intervent, we may refer back to the performance of the MP algorithm with a restricted domain of application, given by the four finest resolution levels of the WP table, among which the energy is distributed according to Table 2.

Figure 4 reports the absolute and the squared ACFs for the residuals from the WP table. We observe that with the WP table the dependence left in the ACF plots is less evident than before, particularly with regard to the long memory component, while the initial autocorrelation decreases with T, thus suggesting that the feature detection power improves qualitatively by simply concentrating the MP activity only on the finest resolution levels.

Table 2. Energy percentage distribution among the 3 finest resolution levels for residual 5 min series obtained in WP table and computed *via* the MP algorithm at the approximation power of 50, 100, 200 and 500 atoms.

T = # of Atoms	50	100	200	500
WP table				
level 0	0.228	0.268	0.339	0.472
level 1	0.139	0.088	0.135	0.120
level 2	0.1	0.146	0.125	0.126
level 3	0.533	0.497	0.401	0.282

6.2 Interpreting the results

The advantages of working with band-pass wavelet filtered detail signals in terms of temporal aggregation effects are known to come from more stationarized and decorrelated signals, being them almost uncorrelated along individual scales and almost independent across scales. In such a non-Gaussian and non-stationary setting, there is still residual time non-homogeneity, due to heteroscedasticity, too. However, the non-Gaussian probabilistic nature of the resolution-wise sequences obtained from the WP-transformed return series is such that ICA performs well.

The MP algorithm benefits from working with least dependent coordinates, since it learns in a faster and better way; the selection of the MRA signals reflects the decomposition provided by ICA on the wavelet expansion coefficients, and the least dependent components lead to more orthogonalized MP and thus an increased efficiency.

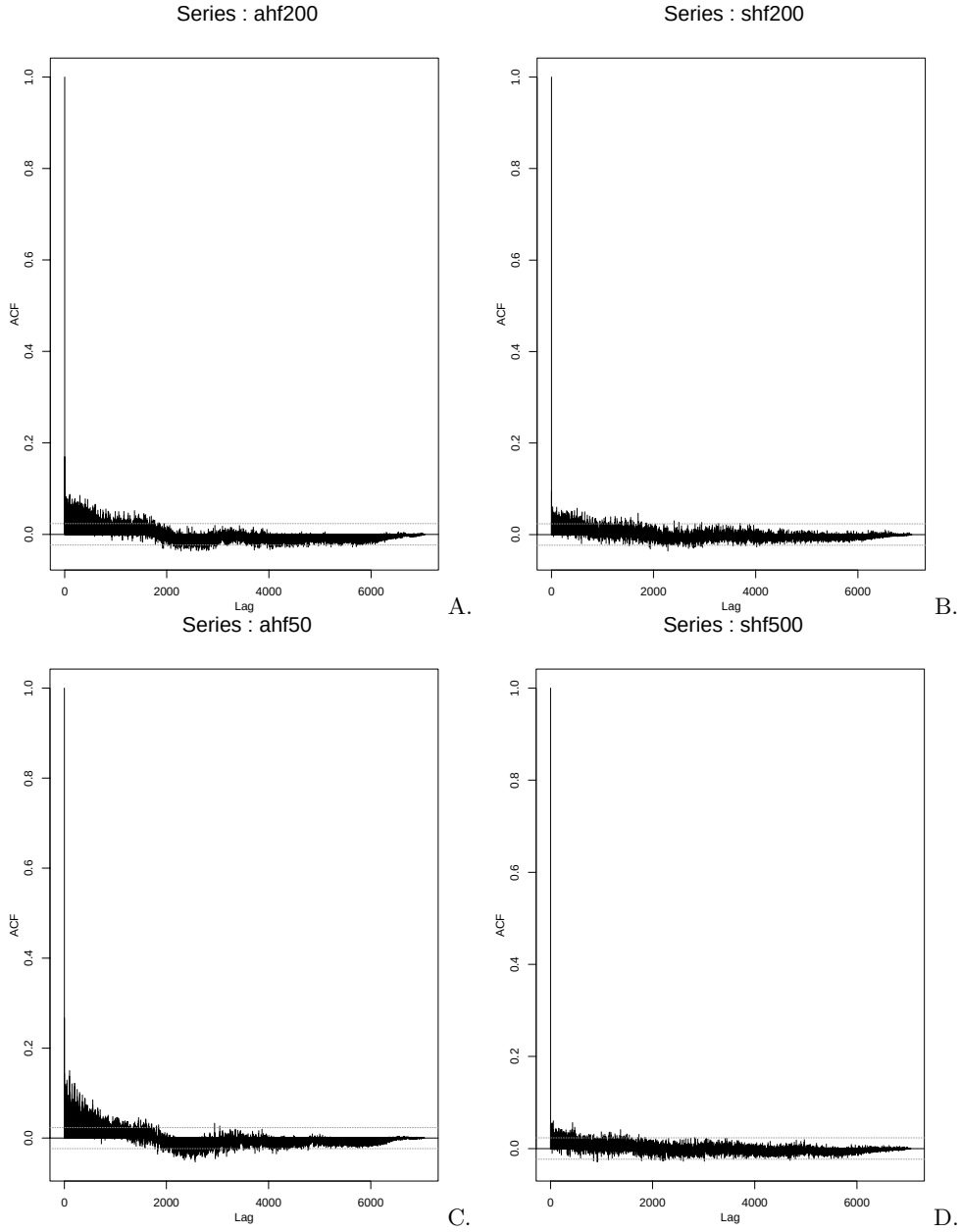

Fig. 4. ACF of 5 min residuals from MP with 200 and 500 atoms on the WP table (A-C are absolute values; B-D are squared values) with the finest four resolution levels.

7 Conclusions

The methodological novelty of this work refers to the possibility of designing a sparse approximation method which enables signal sources separation through an Independent Component Analysis sequentially applied to WP-filtered signals, with the result that a sort of Sparse Component Analysis is obtained.

With high frequency financial time series the results are promising since for the volatility process related to a stock index return index we found a sparse representation together with near optimal decomposition in a set of least dependent components.

The selection of finer resolution levels eliminates redundant information by keeping highly localized time resolution power without simultaneously losing too much frequency resolution; this is due to the fact that low scale information can be reproduced by averaging information from the higher scales.

Sparsity basically means that the covariance matrices have fast off-diagonal elements decay when a locally stationary process is observed. This sparse matrix should be estimated and ideally might be assumed to be a band or near diagonal matrix; one solution is Best Orthogonal Basis, but we have seen that for our time series is sub-optimal compared to the greedy Matching Pursuit, which is found to deliver a near-optimal method of tuning the resolution pursuit.

The author is a recipient of the 2001/02 *ERCIM Research Fellowship* and would like to thank the Nikko Investment Technology Research group formerly based in Los Altos, CA, for the analysis of the data sets.

References

1. P. Abry, P. Flandrin, M.S. Taqqu, D. Veitch, *Wavelets for the analysis, estimation and synthesis of scaling data.* in edited by C. Park, W. Willinger, *Self-similar network traffic and performance evaluation* (Wiley, 2000), pp. 39–88.
2. S. Amari, Tech. Rep., RIKEN, JP (1998).
3. T. Andersen, T. Bollerslev, J. Empirical Finance **4**, 115 (1997).
4. T. Andersen, T. Bollerslev, J. Finance **LII(3)**, 975 (1997).
5. A. Bruce, H.V. Gao, *S+Wavelets*, Seattle: StaSci Division, MathSoft Inc (1994).
6. E. Capobianco, *Wavelets for High Frequency Financial Time Series*, in: *Interface '99 Conference Proc. (1999), pp. 373–378.*
7. E. Capobianco, *Indipendent Multiresolution Component Analysis and Matching Pursuit. Tech. Rep.*, CWI, PNA-R0111 (2001).
8. J. Cardoso, Source separation using higher order moments, in: *Proc. International Conference on Acoustic, Speech and Signal Processing, (1989), pp. 2109–2112.*
9. J. Cardoso, A. Souloumiac, IEE Proc. F. **140**, 771 (1993).
10. S. Chen, D. Donoho, M.A. Saunders, SIAM Rev. **43**, 129 (2001).
11. R. Coifman, V. Wickerhauser, IEEE Tr. Inf. Theory **38**, 713 (1992).
12. P. Comon, Signal Proc. **36**, 287 (1994).
13. I. Daubechies, *Ten Lectures on wavelets* (Philadelphia: SIAM, 1992).
14. D. Donoho, Constructive Approximation **17**, 353 (2001).
15. D. Donoho, I.M Johnstone, Biometrika 81, 425 (1994).
16. D. Donoho, I.M Johnstone, J. American Statist. Assoc. **90**, 1200 (1995).
17. D. Donoho, I.M Johnstone, Ann. Stat. **26**, 879 (1998).
18. D. Donoho, S. Mallat, R. von Sachs, *Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods*, Tech. Rep. (Stanford University, CA US, 1998), p. 517.
19. F. Girosi, Neural Comput. **10**, 1455 (1998).
20. W. Greblicki, M. Pawlak, IEEE Trans. Automatic Control, **31**, 74 (1986).
21. W. Greblicki, M. Pawlak, IEEE Trans. Information Theory **35**, 409 (1989).
22. W. Hardle, G. Kerkyacharian, D. Picard, A. Tsybakov, *Wavelets, Approximation, and Statistical Applications* (Springer-Verlag, New York, 1998).
23. Z. Hasiewicz, Signal Proc. **81**, 791 (2001).
24. A. Hyvarinen, E. Oja, Neural Comput. **9**, 1483 (1997).
25. A. Hyvarinen, *Independent Component Analysis for Time dependent Stochastic Processes*, in: *ICANN'98 Proc. (1998), pp. 541–546.*
26. A. Hyvarinen, IEEE Trans. Neural Networks **10**, 626 (1999).
27. I.M. Johnstone, B.W. Silverman, J. Roy. Stat. Soc., Ser. B. **59**, 319 (1997).
28. H. Krim, J.C. Pesquet, *On the statistics of Best Bases criteria.* in, *Wavelets and Statistics*, edited by A. Antoniadis, G. Oppenheim (Springer-Verlag, New-York, 1995), p. 193.
29. A. Krzyzak, IEEE Trans. Information Theory **36**, 141 (1990).
30. M.S. Lewicki, T.J. Sejnowski, Neural Comput. **12**, 337 (2000).
31. S. Mallat, Z. Zhang, IEEE Tr. Signal Proc. **41**, 3397 (1993).
32. I. Meyer, *Wavelets: algorithms and applications* (SIAM, Philadelphia, 1993).
33. T. Mikosch, C. Starica, Ann. Stat. **28**, 1427 (2000).
34. A. Teolis, *Computational signal processing with wavelets* (Birkhauser, Basel, 1998).
35. M. Zibulewsky, B.A. Pearlmutter, Neural Comput. **13**, 863 2001.