Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Independent multiresolution component analysis and matching pursuit

E. Capobianco

Probability, Networks and Algorithms (PNA)

PNA-R0111 July 31, 2001

Report PNA-R0111
ISSN 1386-3711

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Independent Multiresolution Component Analysis and Matching Pursuit

E. Capobianco
CWI
P.O. Box 94079, 1090 GB Amsterdam
The Netherlands

ABSTRACT
We show that decomposing a class of signals with overcomplete dictionaries of functions and combining multiresolution and independent component analysis allow for feature detection in complex non-stationary high frequency time series. Computational learning techniques are then designed through the Matching Pursuit algorithm, whose performance is monitored so to extract relevant information about the structure of the volatility function. We refer to wavelet and cosine packet dictionaries due to the fact that with intra-daily time series some features of the underlying stochastic processes may remain undetected when standard volatility models are applied to the observed data. Independent component analysis results are particularly encouraging and suggest a better compromise between time and frequency resolutions, and thus a more efficient and accurate Matching Pursuit performance.

2000 Mathematics Subject Classification: 60H30, 62M10, 62G07.
Keywords and Phrases: Latent Variable Systems; Overcomplete Dictionaries; Multiresolution Analysis; Independent and Sparse Component Analysis; Matching Pursuit; Feature Detection; Financial Time Series.

Note: This work was supported by ERCIM.

1. Introduction

In this work we study latent variable systems endowed with complex dynamics, non-gaussian and non-stationary behavior. One of the most recent directions of research in various disciplines has been that of finding relevant information from sparsely represented signals (Donoho, 1996; Lewicki & Sejnowski, 2000; Zibulevsky & Pearlmutter, 2001; Zibulevsky & Zeevi, 2001). Sparse signals require that a small number of expansion coefficients represent them so that the reconstruction quality can be near optimal given the achieved compression power.
Sparsity reminds and refers to statistical parsimony in model building, as opposed to the redundancy of information; one thus may aim to consider and exploit sparsity for inference or signal compression purposes.
Smoothness too is a related concept, both from the standpoint of a function space of objects or signals and from the perspective of a sequence space of expansion coefficients; with a variable degree of smoothness, models are required to be both flexible, in terms of assumptions about probability distributions involved, thus resulting non-parametric, and

adaptive, for dealing well with inhomogeneities of the series.
Given a stochastic process whose realizations might be represented through a certain functional expansion, the idea of decomposing the observed structures in more statistically independent components may be a key goal in applications. Depending on the nature of the process, a more realistic objective could be a search for least dependent components, which sometimes is proposed for dealing with strong forms of dependence and non-stationarity. Wavelets indeed can play an important role in these last cases, since they yield (Johnstone & Silverman, 1997; Abry, Flandrin, Takku & Veitch, 2000) de-correlating and stationarizing effects on the computed coefficient sequences; thus, statistical inference can be more effective in this projected domain.
With regard to financial time series analysis, in previous work (Capobianco, 1999) some wavelet-based methodologies have been proposed and interesting empirical modelling results have been obtained with relevance for the structure and correlation aspects of volatility. In particular, algorithms like the *Matching Pursuit* (MP) have been seen as effectively detecting features in high frequency financial time series.
Here we show that *Independent Component Analysis* (ICA) (Cardoso, 1989; Comon, 1994), or Blind Source Separation (Jutten & Herault, 1991), might be very conveniently adopted in combination with the MP so to artificially learn the structure of a complex class of signals.
We thus suggest a possible way to employ the bank of sources offered by the decomposed signals obtained at different resolution levels from the employed transforms, where each level may give information on market activity with respect to various degrees of temporally aggregated trading horizons. A least dependent component analysis by ICA may thus be combined with the sparsity of signal representation, achieved through *wavelet packet* and *cosine packet* transforms (WPT and CPT, respectively) and related thresholding estimation.
Source separation occurs in the sparse expansion coefficients domain and the signal is reconstructed from resolution levels selected as least dependent ones.
We represent volatility within the frame of latent variable systems where a *Sparse Component Analysis* (SCA) (Donoho, 2000) can be implemented, as suggested by modern signal processing and computational statistics techniques. By pursuing this approach we aim to formulate an initial proposal for innovative views of volatility models.
The paper is organized as follows. Section 2 presents the frame for our modelling approach. Section 3 introduces computational learning issues through wavelet-based techniques and overcomplete dictionaries of functions. Sparsity is addressed together with de-noising and non-linear estimation issues; some optimization algorithms are then described. Section 4 describes ICA and SCA concepts. Section 5 proposes a learning algorithm aimed to improve the time and frequency resolution trade-off. Section 6 reports an experimental analysis based on the approximation of the latent features of the volatility function characterizing a stock returns index. Section 6 concludes the paper.

2. Latent Variable Systems

We start by casting the processes of interest in a very general frame so to represent their dynamics; we thus describe the following linear system:

$$Y_t = A_t X_t + \epsilon_t \tag{2.1}$$

$$X_t = C_t \Phi_t + \eta_t \tag{2.2}$$

where Y_t are observed financial returns[1], 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 $v_t = \sigma_t^2$ can be considered the *volatility process*, which in financial volatility models represents a latent process underlying the returns dynamics.
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* (Olshausen & Field, 1997; Lewicki & Sejnowski, 2000; Chen, Donoho & Saunders, 2001) for the signal under investigation. 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[2].
As far as concerns applications, such system is specialized to the case of studying financial volatility in this work; nevertheless, it can be applied to other different contexts, as shown in other studies (Kisilev, Zibulevsky, Zeevi & Pearlmutter, 2000). Therefore, it may hold as a quite general frame and thus suggests a sort of model-free approach for representing the dynamics of the system of interest.
A special case (Zibulevsky & Pearlmutter, 2001) 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 $Y_t = A_tC_t + A_t\tilde{\eta}_t + \epsilon_t$ or also $Y_t = A_tC_t + \xi_t$, 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:

$$\tilde{X}_t = C_t + \tilde{\eta}_t \tag{2.3}$$

$$Y_t = A_tC_t + \xi_t \tag{2.4}$$

If the *signal-to-noise ratio* (S/N) results high with regard to the sources stochastic nature, 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 volatility becomes characterized by $\Sigma_t = D_t + \sigma_{\epsilon,t}$, where $D_t = A_t\sigma_{\eta,t}\Psi_t + \sigma_{\epsilon,t}$. In the latter case, i.e. when an overcomplete dictionary is available, the estimation procedure of the time inhomogeneous covariance matrix will be conducted through computational learning tools which refer to different techniques, and thus represent an hybrid methodology.
As a result, we have the system (1-2) representing a volatility process; in this way we might generalize the typical autoregressive form of dependence, depending on the structure of the Φ_t matrix[3]. We have the volatility structure expressed non-parametrically and investigated by selected dictionaries of functions, wavelet packets (WP) and localized cosines or cosine packets (CP).
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.
As an alternative frame, we have the system (3-4), where the mixing A_t is now acting on the computed transform expansion coefficients C_t. In other words, one can work in a

[1]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.

[2]Thus the fact that we don't require positivity means that we are not describing volatility through equation (2), but simply sources of it.

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

signal or sequence space, of functions or coefficients, respectively, depending on criteria such as sparsity of representation and statistical independence of the coordinates.
Since the sources are unobservable, estimating them and the mixing matrix is quite complicated; we can either build an optimization system with a regularized objective function through some smoothness priors, so to estimate the parameters involved, or we can proceed more recursively in the mean square sense, through iterations of the MP processing the observed returns with the WP and CP libraries, and looking at $Y_t \approx P_t\Phi_t + \xi_t = A_tC_t\Phi_t + \xi_t$, where the noise is including an approximation error from the system equation and residual measurement effects ϵ_t.
The MP algorithm works on a sparse P_t by the means of overcomplete representations and 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.
Thus, if A_t accounts for modulating the dependence structure of the latent volatility sources, the packet expansion coefficients become the inputs for the ICA step that follows.
The nature of the resolution-wise detail time series is such that ICA naturally fits well, since the series result non-Gaussian and stationary, in the projected sequence space of detail signals too; they are indeed stationarized, as an effect of the wavelet packet transform. There is still inhomogeneity at the detail levels, since they maintain heteroscedastic and thus time-varying features, but this last aspect can be controlled in part by the means of an underlying semi-stationarity hypothesis holding for a segmented version of the initial return process.
With a complete dictionary operating in the new system and obtained by changing the basis allows for the same optimization criteria to apply as well, and one may thus prefer to work with it, i.e. in these new coordinates. From our perspective, the coefficients are now sparsely represented and investigated in separated sources of volatility information through ICA; the original returns have a new decomposition through (4), where the operator A_t enters directly the system dynamics and the sources have changed in (3) from the initial latent volatility components to the transformed and scaled volatilities, embedded in detail signals. The compression and decorrelation properties of wavelet transforms can be better supported with a more effective search for least dependent components via ICA. Our experiments with high frequency financial time series suggest that very good results are obtained through the MP procedure based on WP and CP decomposition dictionaries. Return data may be analysed in two steps, where the first one is a filtering procedure removing all the hidden periodicities, and thus de-seasonalizing the volatility process. The WPT and CPT deliver decomposition tables where one observes how the information is distributed among high and low frequency components, and form the ground for the Matching Pursuit algorithm runs.
The second step is played by ICA which finds what resolution levels appear to have informative content, based on the independent contribution coming from each detail signal to the global signal structure. The MP algorithm yields residuals with autocorrelation and long memory structure, i.e. short and long range dependencies; with ICA these features may thus result more usefully separated from the pure volatility process, which can then be handled with ad hoc de-volatilization models.

3. Computational Learning

3.1 Wavelets and Multiresolution Analysis

Given a *scaling function* or *father wavelet* ϕ, such that its dilates and translates constitute orthonormal bases for all the V_j subspaces obtained as scaled versions of the subspace V_0 to which ϕ belongs, and given a *mother wavelet* ψ together with the terms indicated with ψ_{jk} and generated by j-dilations and k-translations, such that $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$, we obtain differences among approximations computed at successively coarser resolution levels and can form (Daubechies, 1992) a *Multiresolution Analysis* (MRA), i.e. *a sequence of closed subspaces*[4] *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 $L_2[0,1]$.

Signal decompositions with the MRA property have also near-optimal properties in a quite wide range of inhomogeneous function spaces (Daubechies, 1992; Meyer, 1993; Hardle, Kerkyacharian, Picard & Tsybakov, 1998). 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, even if not necessarily orthogonal, are obtained. The scaled and shifted functions $\phi_{jk}(t)$ are Riesz bases for the scaling spaces V_j. On these spaces the signal 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 $D_j(t) = P_{W_j}X(t) = \sum_k d_x(j,k)\psi_{j,k}(t)$, with W_j the wavelet subspace.

The de-correlation effect of the wavelet coefficients is one of the main properties that wavelet transforms bring in the analysis (Johnstone & Silverman, 1997; Abry, Veitch & Flandrin, 1998; Johnstone, 1999). Wavelets characterize function spaces[5], as stated in Daubechies (1992, §9.2, *pp.*298), *"since the* ψ_{jk} *constitute an unconditional basis for* $L^p(R)$*, there exists a characterization for functions* $f \in L^p(R)$ *using only the absolute values of the wavelet coefficients of* f*"*, thus becoming $|< f, \psi_{jk} >|$ the term to look at so to decide whether $f \in L^p$.

From Donoho (1996, §4, *pp.*390), *"when an orthogonal basis is an unconditional basis for a function space F, it means that there is an equivalent norm for the space,* $\|f\|_F$*, such that the ball* $\mathcal{F}(C) = \{: \|f\|_F \leq C\}$ *corresponds to a set of coefficient sequence* $\Theta(C) = \{\theta(f) : f \in \mathcal{F}(C)\}$ *which is solid and orthosymmetric"*, which means that if $\theta \in \Theta$ and $\mid \theta_i^{'} \mid \leq \mid \theta_i \mid, \forall i$, then $\theta^{'} \in \Theta$. As a consequence, an unconditional basis diagonalizes a functional class and retains optimal sparsity.

Generally speaking, with a *Discrete Wavelet Transform* (DWT) a map $f \to w$ from the signal domain to the wavelet coefficient domain is obtained, i.e. one applies, through a bank of *quadrature mirror filters*, the transformation $w = Wf$, so to get the coefficients for high scales (high frequency information) and for low scales (low frequency information). A sequence of smoothed signals and of details giving information at finer resolution levels is found from the wavelet signal decomposition and may be used to represent a signal expansion:

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

[5] This same property can be extended to many function spaces, i.e. Sobolev, Holder, for instance, and in general all Besov and Triebel spaces.

$$f(x) = \sum_k c_{j0,k}\phi_{j0,k}(x) + \sum_{j>j0}\sum_k d_{j,k}\psi_{j,k}(x) \tag{3.1}$$

where $\phi_{j0,k}$ is associated with the corresponding coarse resolution coefficients $c_{j0,k}$ and $d_{j,k}$ are the detail coefficients, i.e. $c_{j,k} = \int f(x)\phi_{jk}(x)dx$ and $d_{jk} = \int f(x)\psi_{jk}(x)dx$. In short, the first term of the right hand side of (3) is the projection of f onto the coarse approximating space V_{j0} while the second term represents the cumulated details. We may define empirical estimates $\hat{c}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\phi_{j,k}(x_i)$ and $\hat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(x_i)$ and consider the advantages of an orthogonal wavelet expansion, which under standard normality assumptions implies finding independent coordinates in the decomposition domain of wavelet expansion coefficients, even in the presence of correlation.

3.2 Wavelet De-noising

In the wavelet-based representations of signals sparsity inspires strategies that eliminate redundant information, not distinguishable from noise; 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* (Donoho & Johnstone, 1994, 1995, 1998) 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.

Given that a better reconstruction might be crucial for financial time series in order to capture the underlying volatility structure and hidden dependence, de-noising can be usefully employed for these spatially heterogeneous signals. The following well-known algorithm is usually implemented:

- *The wavelet transform is applied to the data, so to get empirical wavelet coefficients;*
- *The empirical wavelet coefficients are shrunken toward zero by setting a thresholding rule reflecting the nature of the data and by using suitable and possibly optimal statistical estimation criteria;*
- *The inverse DWT is applied to the thresholded coefficients so to reconstruct the signal in a sparse way.*

The *shrinkage rule* and the *threshold value* are selected among several possible choices, and given the noisy nature of observed financial time series, an *adaptive procedure* might be preferred.

The *soft shrinkage* rule selected is $\delta_s(\tilde{d}_{jk},\lambda) = sgn(\tilde{d}_{jk})(\mid \tilde{d}_{jk} \mid -\lambda)_+$, when $\mid \tilde{d}_{jk} \mid > \lambda$, or otherwise $\delta_s(\tilde{d}_{jk},\lambda) = 0$. It thus keeps or shrinks values, compared to the keep-or-kill solution offered by the *hard rule*, where $\delta_h(\tilde{d}_{jk},\lambda) = \tilde{d}_{jk}I(\mid \tilde{d}_{jk} \mid \geq \lambda)$.

Inhomogeneous function classes characterization, diagonalization and sparsity thus yield, together with the multiresolution property, a powerful justification for selecting wavelets as an approximation and estimation instrument. In representing a function belonging to a general space, space-time resolution combined with frequency resolution are pursued by respectively using contracted (high frequency) and dilated (low frequency) versions of wavelets. Therefore, an increased localization power yields advantages in terms of spatial adaptivity, which might be very useful for handling financial time series.

3.3 Overcomplete Dictionaries

Function dictionaries are collections of parameterized waveforms (Chen, Donoho & Saunders, 2001); they are available for many classes of functions, formed directly from a particular family, like wavelets, or from merging two or more dictionary classes. Particularly in the latter case an overcomplete dictionary is composed, with linear combinations of elements that may serve to represent remaining dictionary structures, thus originating a non-unique signal decomposition.

An example of overcomplete representations is offered by WPs, which represent an extension of the wavelet transform to a richer class of building block functions and allow for a better adaptation due to an oscillation index f related to a periodic behaviour in the series which delivers a richer combination of functions.

Given the admissibility condition $\int_{-\infty}^{+\infty} W_0(t)dt = 1$, $\forall (j,k) \in Z^2$ we have from (Krim & Pesquet, 1995):

$$2^{-\frac{1}{2}} W_{2f}(\frac{t}{2} - k) = \sum_{i=-\infty}^{\infty} h_{i-2k} W_f(t-i) \tag{3.2}$$

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

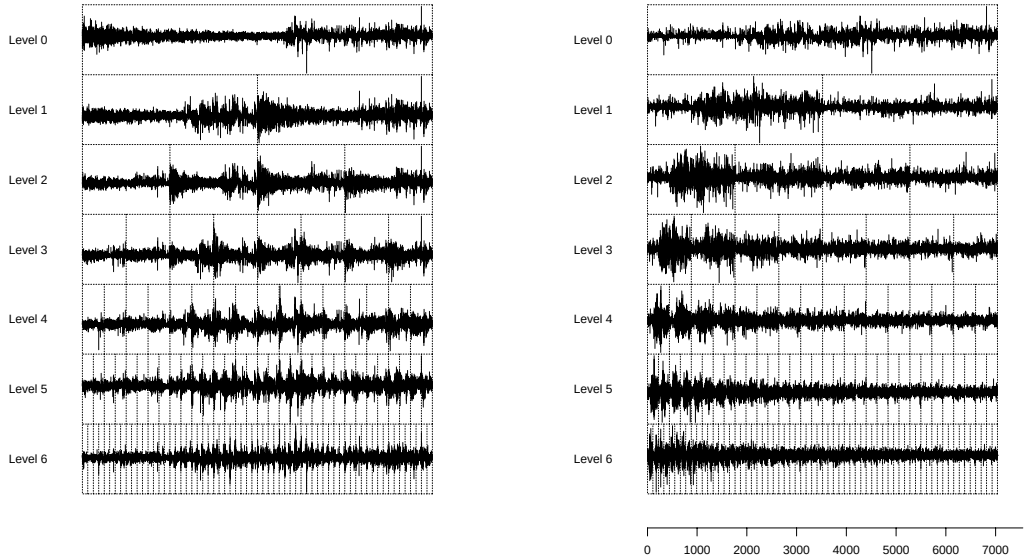
$$2^{-\frac{1}{2}} W_{2f+1}(\frac{t}{2} - k) = \sum_{n=-\infty}^{\infty} g_{n-2k} W_f(t-n) \tag{3.3}$$

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 and an orthogonal wavelet transform can always be selected by changing the partition P and defining $w_0 = \phi(t)$ and $W_f = \psi$, from the so-called WPT we can thus choose combinations of wavelets and other functions reflecting the features of the signal at hand, or search the best basis able to represent the signal with particular sub-sets of coefficients.

The WP representation generalizes other periodic models, like (Li & Xie, 1997) where $y(t) = \sum_{k=1}^{q} \alpha_k exp(it\lambda_k) + \xi(t)$, with $\xi(t)$ a stationary zero-mean time series, α_k random variables uncorrelated to each other and w.r.t. $\xi(t)$, and λ_k the q unknown hidden periodic components. We need to specify a stochastic or probabilistic version of $f(t)$ and allow for the systematic terms to represent the sum of the periodic components of the model, where α_k are the packet coefficients and the exponentials are the dictionary atoms.

With a CPT system we have instead excellent bases as far as concerns compression power, as shown by (Donoho, Mallat & von Sachs, 1996 and 1998), thus getting sparsity of representations through them. Furthermore, in (Mallat, Papanicolaou & Zhang, 1998) CP are shown to be optimal bases for dealing with non-stationary processes with time-varying covariance operators. The building blocks in CP are localized cosine functions, i.e. localized in time and forming smooth basis functions. They are almost eigenvectors of *locally stationary processes*, and thus constitute almost diagonal operators used to approximate the covariance function.



A. B.

Figure 1: CP table (A) and WP table (B) with signal segmentation level-by-level.

A sequence of stochastic processes $X_{t,T}$, $t = 1, \ldots, T$ is called *locally stationary* if there exists a representation $X_{t,T} = \mu(\frac{t}{T}) + \int_{-\pi}^{\pi} A(\frac{t}{T}, \lambda) exp(iwt) d\xi(w)$ such that generalizes the Cramer representation of stationary stochastic processes, with $\xi(w)$ mean zero, orthonormal increments process on $[-\pi, \pi]$ (Dahlhaus, 1993; Neumann & von Sachs, 1995, §3.2.2., Def. 3.1).
It is common to represent a signal from WP dictionaries as $f(t) = \sum_{jok} w_{j,o,k} W_{j,o,k}(t)$ and of CP ones as $f(t) = \sum_{jok} c_{j,o,k} C_{j,o,k}(t)$.
The CPT has an advantage over the classic *Discrete Cosine Transform* (DCT); the latter defines an orthogonal transformation and thus maps a signal from the time to the frequency domain, but it is not localized in time and thus is not able to adapt well to non-stationary signals.
Depending on the *taper* functions we select, the cosine packets decay to zero within the interval where they are defined and in general determine functions adapted to overcome the limitations of DCT. A *DCT-II* transform is defined as:

$$g_k = \sqrt{\frac{2}{n}} s_k \sum_{i=0}^{n-1} f_{i+1} cos(\frac{(2i+1)k\pi}{2n}) \tag{3.4}$$

for $k = 0, 1, \ldots, n-1$, and scaling factor s_k resulting 11 if $k \neq 0$ or n, and $\frac{1}{\sqrt{2}}$ if $k = 0$ or n.
The within-block coefficients of the WP and CP formulations describe their contribution in representing the signal features under a varying oscillation index. The WP table presents crystals, i.e. sets of coefficients, stored in sequency order, according to increasing oscillation index. The CP table presents instead blocks ordered by time and the coefficients within the blocks are ordered by frequency. Figure 1 describes these properties.

The way these plots should be read and interpreted suggests that in WP tables the blocks are ordered by frequency, and within blocks wavelet coefficients are ordered by

time; thus, the low frequency information in the signal is expected to be concentrated on the left side and the high frequency information on the right side of the table[6]. For CP tables, the high frequency part of the signal is now expected on the left side, while the low frequency behavior appears from the right side.

3.4 The Matching Pursuit learning algorithm

The design of optimal algorithms is strictly dependent on the adoption of adaptive signal approximation techniques, built on sparse representations. Sparsity refers to the possibility of considering only few elements of a dictionary of approximating functions selected among a redundant set. The MP algorithm (Mallat & Zhang, 1993) is a good example, and it 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 and CP tables, whose signal representations are given by:

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

and

$$CP(t) = \sum_{jfk} c_{j,f,k} C_{j,f,k}(t) + res_n(t)$$

This choice offers some advantages, which we summarize as follows:

- the approximating kernels are flexible with regard to the type of functions used, i.e. localized cosine functions and variably oscillating wavelets;
- the mixtures of functions employed work in space/time and scale/frequency dimensions, thus yielding better spatial adaptivity and localization power;
- a priori or signal-dependent knowledge may be accounted for, by selecting indexed functions or by reducing the problem dimension through the use of a restricted sub-set of functions in the analysis.

The Procedure 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 , of functions whose form should ideally adapt to the characteristics of the signal at hand. The MP decomposition exists in orthogonal or redundant version and 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 in case iteratively re-decomposed according to:

$$f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + res_n(t) \tag{3.5}$$

1. inizialize with $res_0(t) = f(t)$, at i=1;
2. compute at each atom H_γ the projection $\mu_{\gamma,i} = \int res_{i-1}(t) H_\gamma(t) dt$;

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

3. find in the dictionary the index with the maximum projection,

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

which equals from the energy conservation equation $argmax_{\gamma \in \Gamma} \mid \mu_{\gamma,i} \mid$;

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 $i \leq n$.

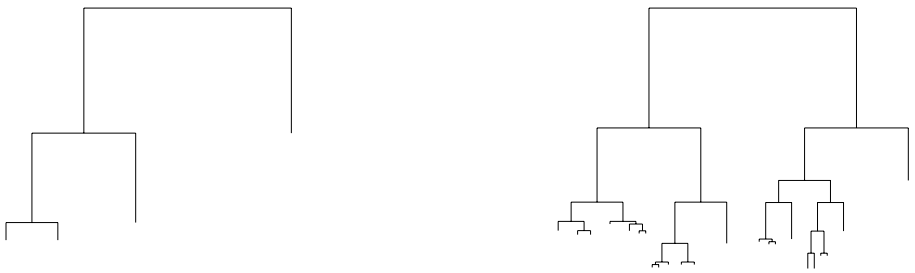
With $\mathcal{H}$ as an Hilbert Space, a function $f \in H$ is decomposed in this frame as $f =< f, g_{\gamma_0} > g_{\gamma_0} + Rf$, with f approximated in the g_{γ_0} direction, orthogonal to Rf, such that $\|f\|^2 =|< f, g_{\gamma_0} >|^2 +\|Rf\|^2$. Thus, the minimization of the $\|Rf\|$ term requires 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 , $_0 \in$, (see Mallat & Zhang, 1993, for further details).

Algorithmic Features and Limitations The main aspect of interest for the computational learning power of the MP algorithm has appeared in our study like in many others, and refers to how is capable of dealing efficiently with the so-called (Davis, Mallat & Avellaneda, 1997), *coherent structures* compared to the *dictionary noise* components. The terminology is used for stressing the importance of learning the most informative structures by the means of the atoms in the dictionaries; this usually happens efficiently at the beginning of the MP operations but only up to a certain iteration time, when the algorithm finds noise structures instead of relevant signal features.
This aspect has been deeply investigated in the mentioned work, and in our application has been controlled by looking at the behaviour of the residue term after n approximation steps; the residue absolute and squared values allow for the autocorrelation functions to give information about the conditional variance, and thus are of direct interest for the volatility modelling aspects.
There is also the risk of learning non-features, or that the algorithm overfits, and thus learns noise (Jaggi, Karl, Mallat & Willsky, 1998). In our case we found that a solution is to modify the range of application of the MP algorithm, thus making it more orthogonalized.
The MP decomposition is nonlinear, but maintains, along its operations:

$$\|R^n f\|^2 = |< R^n f - R^{n+1} f >|^2 + \|R^{n+1} f\|^2 \tag{3.6}$$

an energy conservation law of the following form:

$$\|f\|^2 = \sum_{n=0}^{m-1} |< R^n f, g_{\gamma_n} >|^2 + \|R^m f\|^2 \tag{3.7}$$



A. B.

Figure 2: WPT Best Basis Tree Plot (A), and a correspondent plot for CPT (B).

equivalently as for linear orthogonal decompositions.
There is (Davis, Mallat & Zhang, 1994) a version of MP that selects several vectors from the dictionary at every iteration step, and projects the residue over the space spanned by these vectors; it is an orthogonal MP and for every selected vector computes an orthogonalization step through a Gram-Schmidt algorithm. Despite its faster residue decrease and its convergence in a finite number of steps M, compared to the finite vector dimension N, it presents a relevant implementation cost and also possible numerical instability problems due to ill-conditioning of the basis functions $\{g_{\gamma_n}\}_{0 \leq n < M}$. The strategy we have chosen is explained later and fits with other objectives pursued in the study.

3.5 The Best Basis Algorithm
The *Best Orthogonal Basis* (BOB) algorithm (Coifman & Wickerhauser, 1992) is employed here as an alternative to the MP optimization method, with the goal of minimizing an additive[7] 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 WP or CP dictionaries. The BB 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).
In particular, the BOB steps find a minimum entropy transform from the dictionary at hand, since the above objective function corresponds to $min\ [entr\ f(B)] \mid B \in ,$, where B is an orthobasis in the selected dictionary , 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 $\hat{w}_{j,f,k} = w_{j,f,k} \times (||\ w_{0,0}\ ||_2)^{-1}$. The algorithm is known to deliver near-optimal sparsity representations, but not in the presence of non-orthogonal contexts.
Figure 2 reports the tree plots visualizing the relative entropy content of the packet coefficients, where the arcs represent entropy savings in going from the parent to the child node; the longer ones suggest advantages in adopting the relative wavelet transform.

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

In the WP tree the best basis is concentrated in the three highest resolution levels, indicating an homogeneous entropy reduction among levels. With the CP tree the best basis results much more spread among the resolution levels and shows a superior entropy reduction. 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. $E_j^s + \sum_{j=1}^{J} E_j^d$, where $E_j^s = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^J}} s_{j,k}^2$ and $E_j^d = \frac{1}{E} \sum_{k=1}^{\frac{n}{2^J}} d_{j,k}^2$, for $j = 1, \ldots, J$.
Thus, the CP crystals (i.e. sub-sets of coefficients) are from a wider basis across the resolution levels and form the building blocks selected with a different energy distribution compared to the WP case.
In (Saito, 1998; Saito, Larson & Benichou, 2000) there is a proposal of an alternative view of the BOB scheme with modifications addressing the search for *least statistically dependent bases*. An operator called feature extractor acts for reducing the dimensions of the problem and allows for a change of coordinates, and thus of basis, in the signal domain followed by a selection of m coordinates. The following functional summarizes the scheme, by seeking the best coordinates B^* measuring the efficiency of the bases B spanning $x \in F_m$, given the training set τ and the set of all such bases L, $B^* = argmax_{B \in L} F(B \mid \tau)$.
In Figure 3 we report the top-100 largest coefficients approximation with the BB and the MP algorithms after running on WP and CP dictionaries. We show the BB on the WP table in (A), and on the CP table in (B), while for the MP algorithm we refer respectively to (C) and (D).

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, while the CP cost table suggests that wider ranges of frequencies, including higher frequencies, are captured.
The plots suggest that BB 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 and CP dictionaries, and it does so throughout more scales and by extending the basis which represents the signal.

4. Independent and Sparse Component Analysis

4.1 Searching Independent Components

The goal of searching for statistically independent coordinates characterizing certain objects and signals, or otherwise for least dependent coordinates, due to a strong dependence in the nature of the stochastic processes observed by the structure of the data, leads to ICA or to least dependent best basis algorithms. The combination of these goals with that of searching for sparse signal representations suggests hybrid forms of SCA.
We present the results obtained with ICA, whose role has gained relevance to applications in many fields, particularly signal processing and neural networks. There are still relatively few ICA applications in the domain of finance. Statistically independent components may offer a possible interpretation of the main driving forces behind financial time series, in line with other decomposition techniques such as structural time series analysis or factor models, involving multivariate time series, optimal investment portfolios, component extraction and separation of noise from true prices (Back & Weigend, 1997; Wu & Moody, 1996).

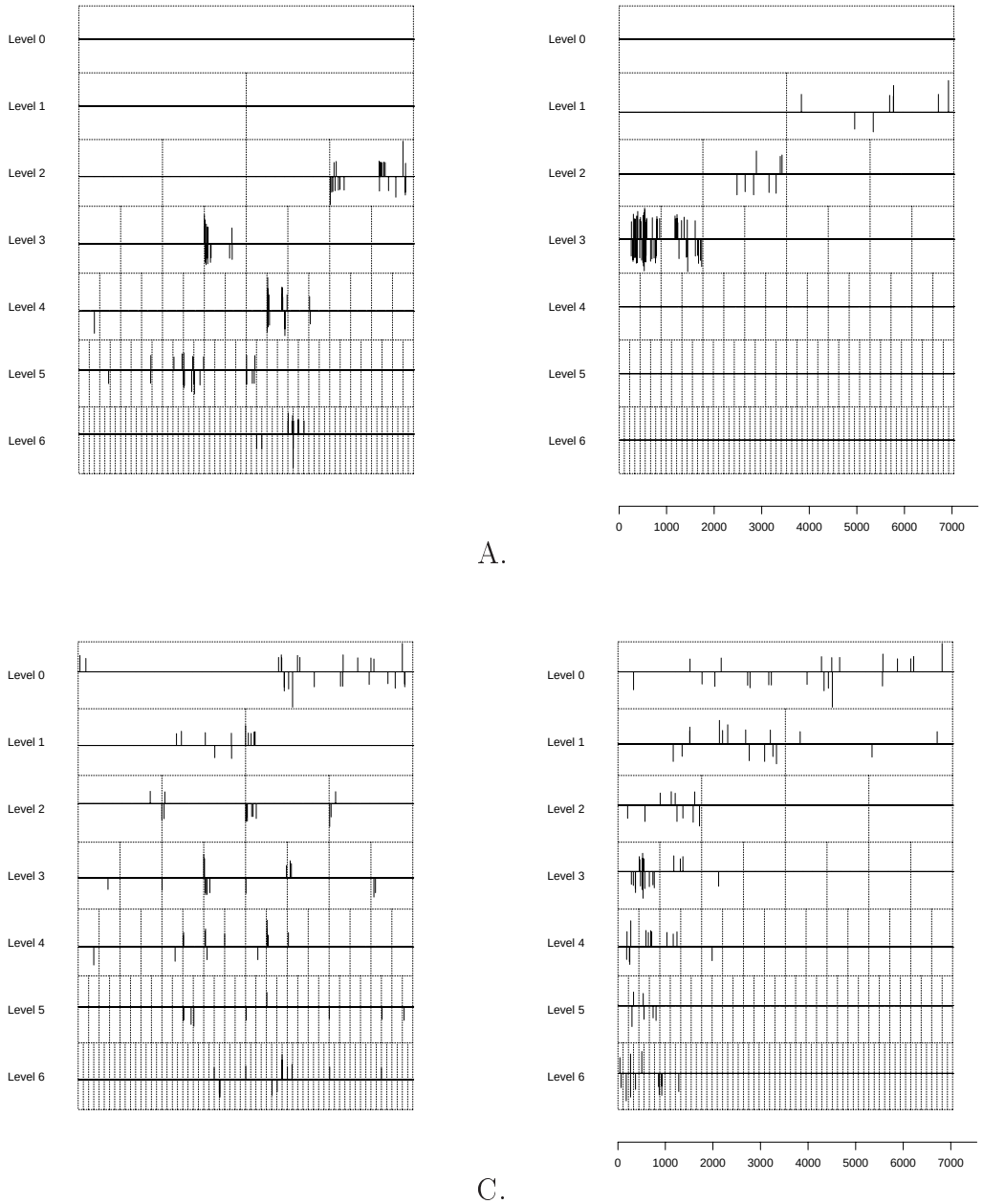


Figure 3: Signal approximation with the 100 largest coefficients for BB run on CP (A) and WP (B), and for MP run on CP (C) and WP (D).

The independent components can be efficiently computed by ad-hoc algorithms such as *JadeR* (Cardoso & Souloumiac, 1993) or *fastICA* (Hyvarinen & Oja, 1997; Hyvarinen, 1999). For Gaussian signals, the *Independent Components* are exactly the known *Principal Components*; with non-Gaussian signals ICA delivers superior performance, due to the fact that it relies on high order statistical independence information.
With SCA one attempts to combine the advantages delivered by sparsity of signal representation, which transfer to better compression power and estimation in minimax sense. Now the expansion coefficients represent sparse vectors, those few large coefficients able to reconstruct the original signal features. The goal is to optimize sparsity so the get optimal recontruction.
ICA is related to linear and nonlinear mixture models, including the case of convolutive mixing, and refers to noise-free or noisy data applications. In studies based on time series, one might find more convenient to work with innovation processes (Hyvarinen, 1998) derived from conditional values of observation processes and usually more independent and non-Gaussian then the original ones.
Likewise, temporal correlation and convolutive mixing (Amari, 1998; Attias, 1998 and 1999) affect the data and their dependence features, such that more elaborated schemes are needed so to account for these aspects too.
As said, ICA generalizes the well-known Principal Component Analysis, but unlike the latter which uses statistical information coming from the first two moments of the involved probability distributions, it decorrelates the data by using statistical information of higher order and thus becomes suitable for non-Gaussian contexts. Therefore, ICA is a latent variable statistical model where linear or non-linear transforms of non-Gaussian and independent variables deliver the observed data.
By assuming that the sensor outputs are indicated by $x_i, i = 1, \ldots, n$ and represent a combination of independent, non-Gaussian and unknown sources $s_i, i = 1, \ldots, m$, a non-linear system $Y = f(X)$ could be approximated by a linear one AS, where $X = AS$. Instead of computing $f(X)$ one may now work for estimating the sources S together with the $m \times m$ mixing matrix A, where usually $m << n$, with n the number of sensor signals, but with $m = n$ holding in many cases too.
The *Joint Approximate Diagonalization of Eigenmatrices for Real signals* (i.e. JadeR) algorithm is the of algorithms implementing ICA that we have applied. It delivers an estimate for the separating or de-mixing matrix B, obtained from $Y = BX$, such that when $B = A^{-1}$ a perfect separation would be obtained. This in general cannot happen, being just an ideal setting, and thus solutions hold approximately up to permutation and scaling. De-correlation and rotation steps are implemented so to deal with these aspects, and a set of approximately m independent components is obtained.
The approach of combining MRA and ICA that we have adopted here is different from other cases of study; wavelet signal decomposition and ICA for financial data analysis has also been suggested by (Wu & Moody, 1996), but with a different goal, i.e. decomposing the stock price series into independent components so to extract the true price from the noisy series. More recent work has been proposed by (Kisilev, Zibulevsky, Zeevi & Pearlmutter , 2000) with applications to musical sounds and images.
We start from considering the detail signals obtained through WP and CP transforms: the series of scaled signals bring a different degree of resolution and refer to specific information obtained by the transforms while switching between resolution levels. Then, we combine an ICA step with the MP algorithm operating on WP and CP tables; through such a joint search for sparsity and statistical independence we are basically adopting

an hybrid SCA solution, since we aim to optimize sparsity through the choice of ad hoc function dictionaries, like localized cosines and orthonormal wavelet bases, and because we adopt thresholding estimators. Furthermore, we want to operate through least dependent coordinates such that an almost diagonal covariance operator is achieved, helping the interpretation of latent volatility features.

Searching sparse decompositions From (Donoho, 1996) functions represented as $f \sim \sum_{i=1}^{\infty} \theta_i \phi_i$ have sparsity in their expansion coefficients θ_i which can be measured by appropriate norms targeted to achieve bounds on the performance of compression and de-noising schemes. In this signal representation, both the coefficients θ_i and the basis components ϕ_j have to be computed; searching for the best basis is combined with the requirement of sparsity.
Dictionaries which are overcomplete deliver non-unique signal decompositions; when instead a basis may be selected, the dictionary will result complete. In our applications the hybrid method we have designed requires that the least dependent resolution levels are to be selected by ICA and used for calibrating the MP algorithm, thus achieving a better detection power for the dependence structure in the series.
More independent coordinates along which to apply the algorithmic steps allow the MP to be more orthogonalized and thus work more efficiently in retrieving the coherent structures; the algorithm learns more effectively, working progressively toward obtaining a final residue whose absolute and squared transforms might reveal only pure volatility features.
Alternative models can be designed, and following (Zibulevsky & Pearlmutter, 2001), the elements A and C can be computed from the following optimization problem:

$$min_{A,C} \frac{1}{2\sigma^2} \mid\mid AC\Phi - X \mid\mid_F^2 + \sum_{j,k} \beta_j h(c_{j,k}) \tag{4.1}$$

or following (Girosi, 1998; Poggio & Girosi, 1998) a connection to *Support Vector Machines* and sparse representations can be made by changing the norm in the previous equation.
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, this functional generalizes other similar structures like the *Method of Frames*, the *Basis Pursuit* or the equivalent *Linear Programming* problem representations, perturbed or not depending from the fact that one is considering a noisy observation system or not (Chen, Donoho & Saunders, 2001).
The term $AC\Phi$ can be replaced, with a number of sensors equal to the number of sources, and the inverse mixing or 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 $\mid\mid C\Phi - BX \mid\mid_F^2$.

5. Approximating Stock Index Volatility

5.1 The General Setting

Wavelet orthogonal bases are unconditional bases for certain classes of functions, generally belonging to inhomogeneous function spaces; as such they represent almost diagonal covariance operators, as shown by (Mallat, Papanicolaou & Zhang, 1998). They deliver optimal de-noising and compression ability (Donoho, Mallat & von Sachs, 1996 and 1998).
Certain classes of processes, like locally stationary processes, address the fact that non-stationarity behaviour occurs in some periods of time depending on the presence of regime

shifts or shocks or even other independent factors, and then a switch to a more stationary regime is observed for the variables of interest.
The dependence structure in high frequency financial time series can be detected by selecting ad hoc function dictionaries, like WP and CP, whose good time-frequency resolution trade-off allows for an excellent representation of non-stationary or time inhomogeneous series. Since the signal transformed in the wavelet coefficient domain is heteroscedastic and non-Gaussian, the same two properties transfer to the expansion coefficient domain too; stationarity and decorrelation take place even if some weak dependence structure remains.
We refer 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 (1m). The total sample has 35,463 observations, with intra- daily trading prices covering the working week, holidays and weekends excluded. We then form a temporally aggregated time series of correspondent five-minute (5m) data from the original one; thus, they are simply given by the average of components sampled at the time interval of one minute. The aggregated sample consists of 7092 observations[8].
Model design tasks involve the representation of features such as short and long range dependence, hidden periodicities, external shocks, surprise variable effects and other factors with impact on prices and returns (Andersen & Bollerslev, 1997). Together with the volatility persistence observed from the absolute and squared returns autocorrelation functions, long range dependence seems a typical feature which is often indicated as present in high frequency financial series. It is very likely that this form of dependence might be mixed with other forms of hidden dependence in the data like, for instance, periodicities (see Figure 4). These last components are usually not easy to interpret, and may prevent the researcher from detecting and evaluating the underlying low frequency dynamics, as also suggested by the presence of non-stationarity through the evidence of spurious features in the data.

We adopt a strategy which aims to pre-process the return series with ad-hoc filtering, i.e. targeted to deal with the hidden periodic components. The goal is that of getting residual returns where the only dynamics left are those strictly related to the volatility process. In practical terms, we have in mind a two-stage process where the battery of wavelet-based techniques and the classes of functions available through the selected dictionaries may enable a de-seasonalization step followed by a de-volatilization step. Here we cope mainly with the former aspect, while the latter should require specific volatility modelling too, not proposed here.

5.2 Non-parametric Estimation

We keep this setting of underlying conditions, and thus consider our setting inherently non-stationary; as such, we want to design a method explicitly accounting for these conditions. Thus, after having segmented appropriately the data we run experiments which are based on the methods already introduced. We have tested the MP approximation power, and let the algorithm work with 50, 100, 200 and 500 atoms from the selected WP and CP function dictionaries, so to verify whether its computed residue might be interpreted as noise and might indicate how efficient is learning.
We apply de-noising to the tables so to let the shrinkage principle operate via thresholding and verify whether the MP performance is influenced by the presence of noisy wavelet crystals. We combine the advantage of using dictionaries which are effective in detecting

[8]The experiments were conducted with *S+Wavelets* (Bruce & Gao, 1994).

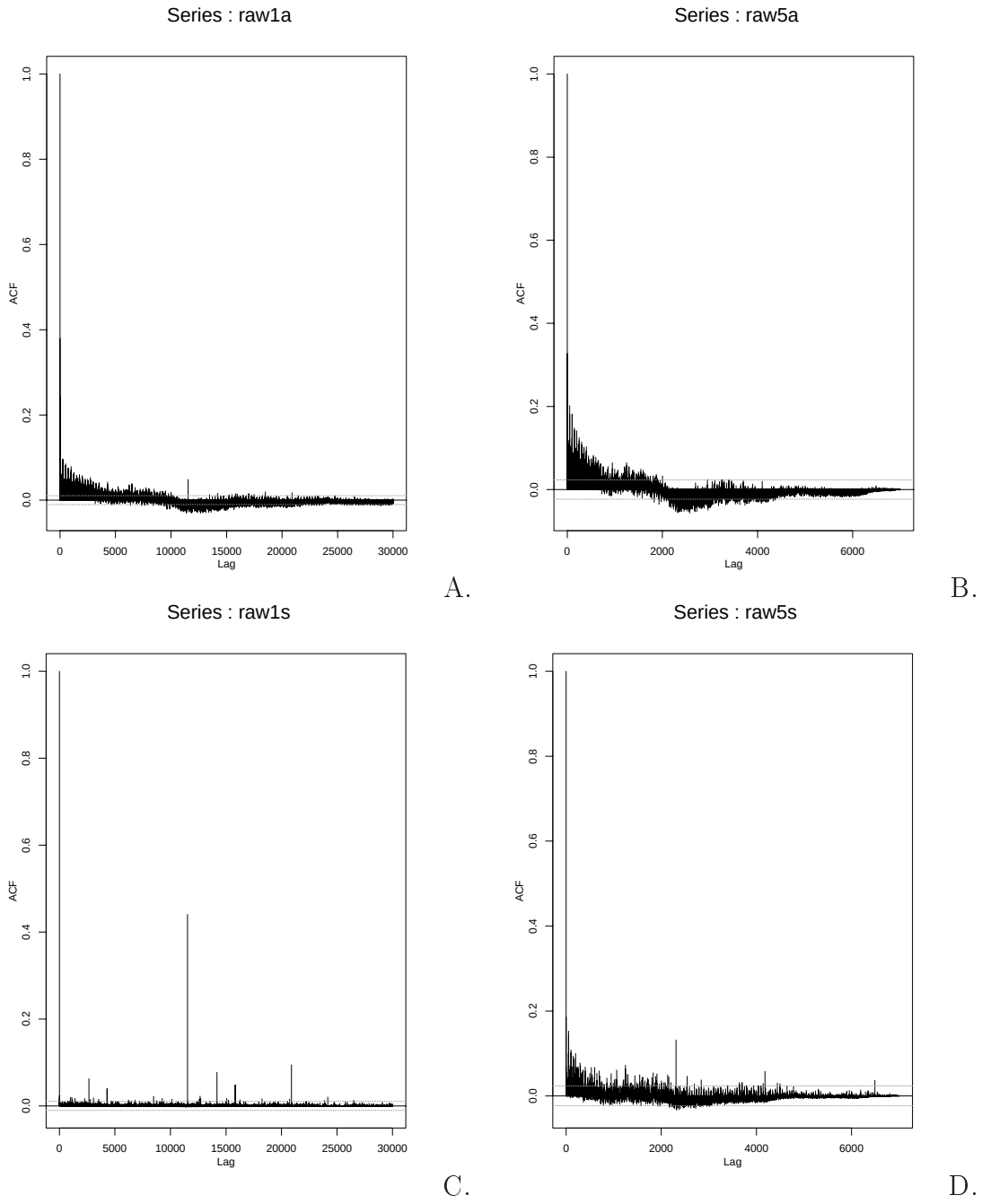


Figure 4: Absolute (indexed by a) and squared (indexed by s) raw 1m and 5m returns.

the latent periodic structure with that of removing the noise characterizing the finest details. In this way we want to reach a better sparsity of coefficients. Thus, the procedure we adopted is described as follows:

- **Step 1**.

  Wavelet/Cosine Packets Segmentation: the initial sample is split into segments. This has been done according to systematic rules, variable according to (Gao, 1997; Mallat, Papanicolaou & Zhang, 1998; Serroukh, Walden & Percival, 2000; von Sachs & MacGibbon, 2000) while here the procedure reflects the sample splitting rule which restricts the partition choice to sample sizes divisible by 2^J in the wavelet/cosine packet analysis.
  We keep the rule at its simplest level, by just considering two sub-samples, with observations ranging from 1 to 3328 and then from 3329 to 7040, for the 5m series. Together with a certain computational advantage, one gets an improved local fit power for estimating the variance by looking more specifically at the data dynamics belonging to less non-stationary segments, which may correspond to separate market phases.

- **Step 2**.

  The Thresholding Algorithm:

  - *The WP and CP transforms are applied to the returns and the empirical wavelet crystals (i.e. sets of coefficients) are computed;*
  - *The empirical coefficients are shrunken toward zero by a thresholding step, which works according to a series of rules reflecting the nature of the data and following optimal statistical estimation criteria;*
  - *The inverse transforms are applied to the thresholded coefficients so to reconstruct the signal in a sparse way.*

  A widely employed threshold which adapts to each resolution level is obtained through the principle of minimizing levelwise the *Stein Unbiased Risk Estimator*, or *SURE*. The resulting estimator is quoted in the literature as *SURE-Shrink*. Therefore one gets:

$$\lambda_j = argmin_{t \geq 0} SURE(d_j, t) \tag{5.1}$$

  and through the following functional:

$$SURE(d_j, t) = K - 2\sum_{k=1}^{K} I_{[|d_{j,k}| \leq t\sigma_j]} + \sum_{k=1}^{K} min[(\frac{d_{j,k}}{\sigma_j})^2, t^2] \tag{5.2}$$

  can find a function estimator like:

$$\hat{f}(x) = \sum_k \hat{c}_{j0}\phi_{j0,k}(x) + \sum_{j>j0}\sum_k sgn(\hat{d}_{j,k})(\mid \hat{d_{j,k}} \mid -\lambda)_+\psi_{j,k}(x) \tag{5.3}$$

The shrinkage function depends also on the estimate of the scale of the noise, which in our application represents a very important aspect. One may use all the coefficients to yield the estimate, or just those ones belonging to each resolution level. A different bias-variance ratio naturally follows in the applied smoothing. We used the estimate from all the crystals, not to lose efficiency and because we rely on a certain dependence structure among resolution levels; thus, we adopted the MAD function, defined by $median(\mid x - median(x) \mid)/0.6745$, which eliminates the noise and delivers a robust variance estimate.

- **Step 3**.

  The MP Algorithm: we apply it to the sub-tables, i.e. to the sampled segments previously computed and run MP with an increasing approximation power, in both the original and the de-noised sub-tables.

- **Step 4**.

  The Energy Distribution: we compare it among series decomposed by resolution levels, and for each sub-table, so to verify which of them are more or less informative and up to what degree the presence of noise gives a contribution to the observed data features. We thus check how the approximation power of the MP algorithm is affected by the noise, and look at the usual diagnostic autocorrelation function (ACF) plots for absolute and squared residuals, which are very informative about the structure of dependence in the volatility process. Ideally, coherent structures should be removed and the algorithm should be stopped when dictionary noise is encountered. When no structure is found in the residue it means that the MP worked efficiently; this fact should also be interpreted as the evidence that only pure volatility aspects are left in the residual series.

We observe from Table 1 that in the first sub-sample of the WP table level 0 increases with T (the number of approximating structures or atoms in the dictionary) and level 6 becomes dominant with de-noised crystals; the latter is followed by level 4, with both the levels decreasing in energy percentage with T, and by level 2, increasing instead with T. The second sub-sample has still level 0, which increases with T, followed by level 3, decreasing with T; this segment concentrates most of the energy from the MP runs on the original noisy WP table, while the waveshrunken crystals indicate level 3 as the one with the largest energy, decreasing with T, followed by level 2, 4 and 6, all pretty much stable in their energy distribution, according to the approximation power employed by the MP algorithm.
In short and as expected, we have observed a shift from fine resolution levels to low and to mid-coarse ones, respectively in the first and second WP table sub-sample, when de-noising is applied through the SURE-SHRINK thresholding (see Figure 5).

In Table 2, in the CP dictionary, we observe that with the original crystals the MP computations suggest levels 0 and 3 together with levels 0 and 1 as dominant, respectively

T = # of Atoms	50	100	200	500	w50	w100	w200	w500
n=1-3328								
level 0	**0.578**	**0.7**	**0.78**	**0.821**	0.0	0.0	0.0	0.0
level 1	**0.139**	**0.155**	**0.116**	**0.1**	0.004	0.011	0.019	0.056
level 2	0.032	0.034	0.026	0.032	0.084	0.192	0.231	0.26
level 3	0.087	0.032	0.03	0.018	**0.105**	**0.079**	**0.112**	**0.123**
level 4	0.038	0.026	0.019	0.015	**0.226**	**0.206**	**0.179**	**0.172**
level 5	0.032	0.017	0.01	0.006	0.04	0.05	0.042	0.048
level 6	0.094	0.037	0.019	0.008	**0.541**	**0.463**	**0.415**	**0.341**
n=3329-7040								
level 0	**0.383**	**0.361**	**0.479**	**0.578**	0.0	0.0	0.0	0.0
level 1	**0.203**	**0.292**	**0.243**	**0.207**	0.012	0.011	0.018	0.026
level 2	0.069	0.096	0.098	0.09	**0.253**	**0.335**	**0.301**	**0.321**
level 3	0.253	0.18	0.118	0.083	**0.532**	**0.425**	**0.398**	**0.338**
level 4	0.062	0.048	0.042	0.027	**0.079**	**0.092**	**0.114**	**0.126**
level 5	0.006	0.007	0.008	0.008	0.022	0.046	0.073	0.078
level 6	0.024	0.016	0.012	0.008	**0.102**	**0.093**	**0.096**	**0.11**

Table 1: Energy percentage distribution among resolution levels for sub-sampled residual 5m series transformed via WPT and computed via MP algorithm at different degrees of approximation power, i.e. with 50, 100, 200 and 500 atoms. MP runs with the original (left part of the table) and de-noised crystals.

in the first and second sub-samples, even if with a different degree of influence of the approximation power employed by MP, while the de-noised crystals suggest that the energy remains pretty much concentrated in the same levels, 0, 1 and 3 in the first segment and 0, 1 and 2 in the second one.
Thus, the finest resolution levels are those with most of the energy and de-noising doesn't really lead to a shift of energy among the scales in the CP tables, compared to the WP tables. This indicates that the CP table is already sparsely representing the signal.

As a final check and so to understand how the approximation power transfers to advantages in feature detection ability, we look at the ACFs computed on the absolute and squared transformed residuals, obtained from the WP/CP tables and their de-noised versions. From the plots in Figure 6 and 7, for the absolute returns ACFs, and in Figure 8 and 9 for the squared values ACFs, we notice that the residual autocorrelation and the persistence remain visible features, particularly with the absolute values, and regardless the approximation power considered, due to either the undetected structure or the algorithm sub-optimal performance (adaptation to non-features, noise overfitting, lack of efficiency as possible causes).
The ACFs computed over the de-noised residuals indicate that with the WP tables these features are less evident while with the CP tables they appear even more emphasized, thus suggesting that the noise, somehow spuriously, contributes to the structure shown by the WP/CP tables. From the squared transforms the effects of de-noising are more visible when looking at the power of detecting the hidden periodicities, since they are strongly highlighted in the WP case and, at a less degree, with the CP tables too.

In summary, the noise seems to hide periodic components and its removal allows for a better detection of them, and this suggests that non-stationarity is very likely responsible for the presence of spurious features. While the CP transform suggests a good low and

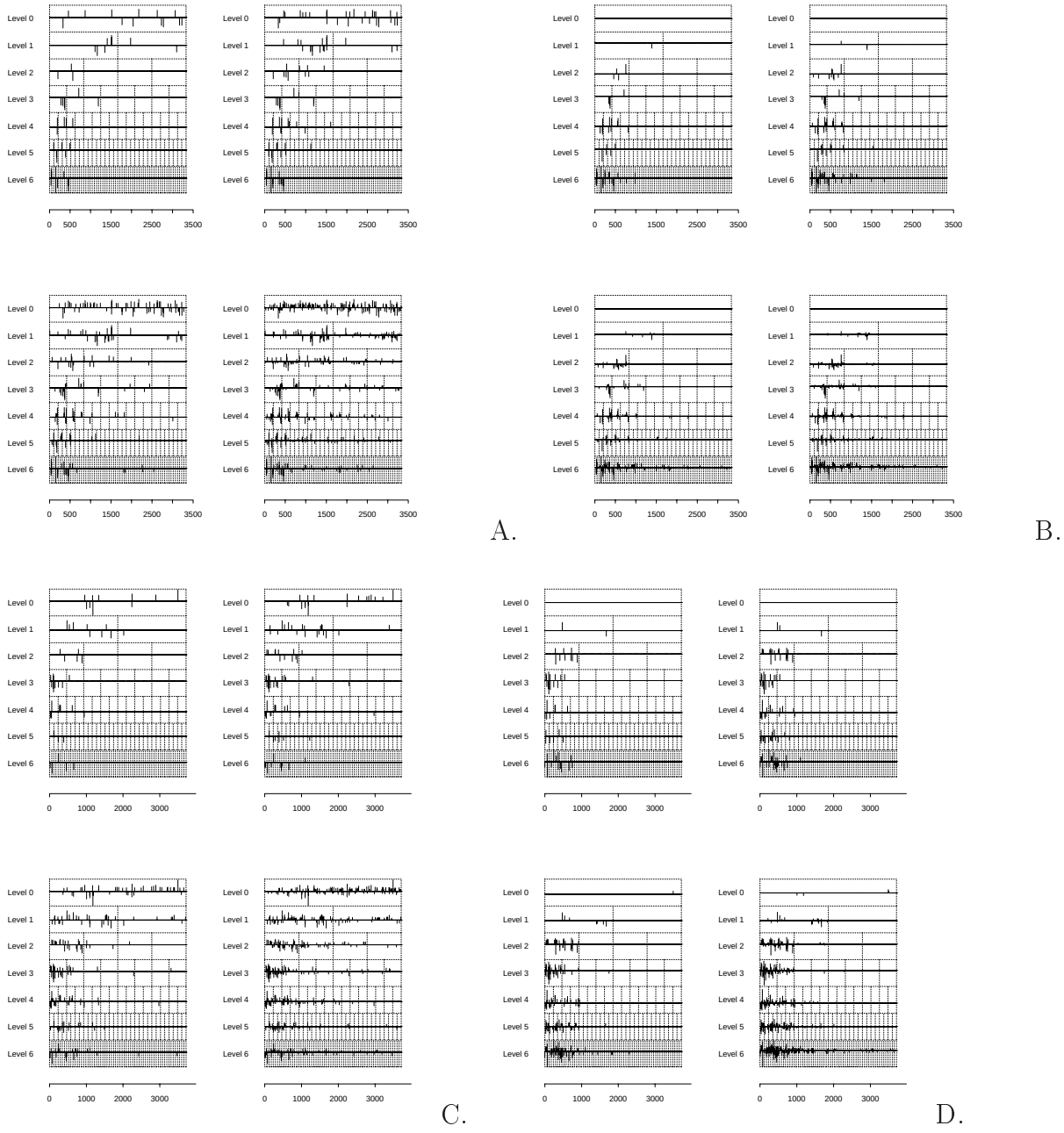


Figure 5: MP progressive approximation power. For each block of plots, 50 (top left), 100 (top right), 200 (bottom left) and 500 atoms used, the first and second sub-samples of the original and de-noised WP tables are indicated, respectively, by A-B (1st segment) and C-D (2nd segment).

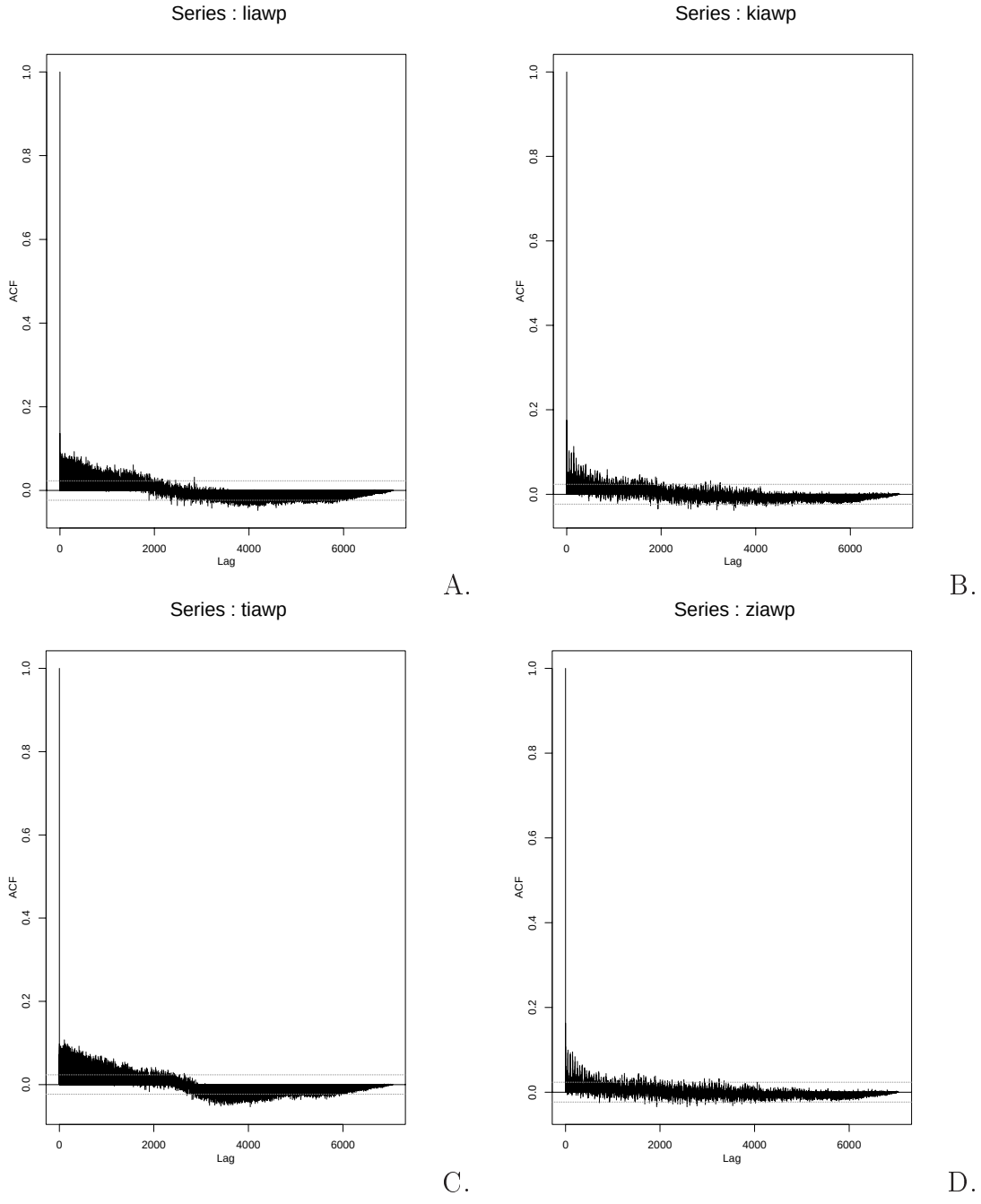


Figure 6: ACF of absolute 5m residuals from MP with 200 (A-B) and 500 (C-D) atoms on the WP dictionary for respectively original (left) and de-noised (right) tables.

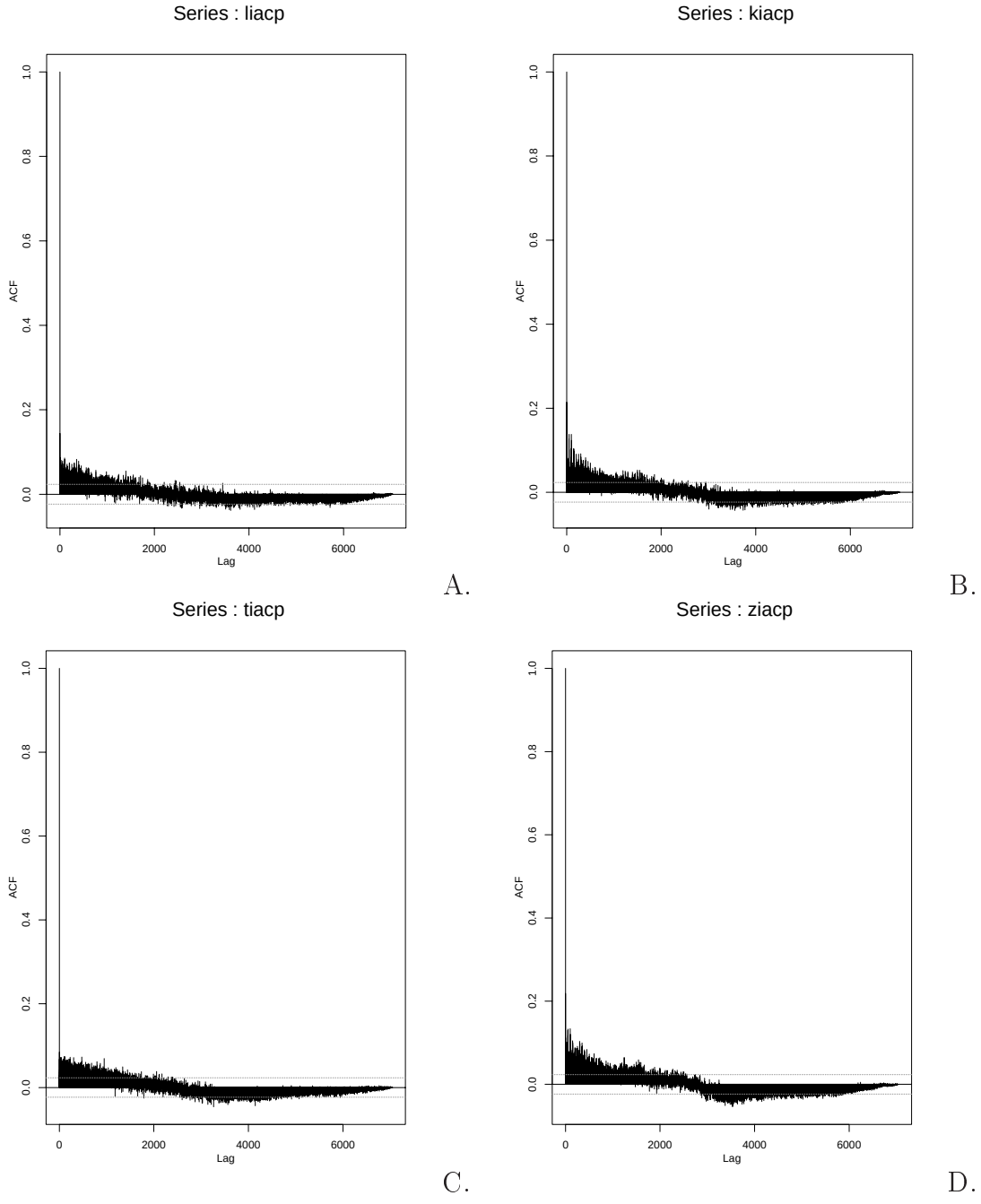


Figure 7: ACF of absolute 5m residuals from MP with 200 (A-B) and 500 (C-D) atoms on the CP dcitionary for respectively original (left) and de-noised (right) tables.

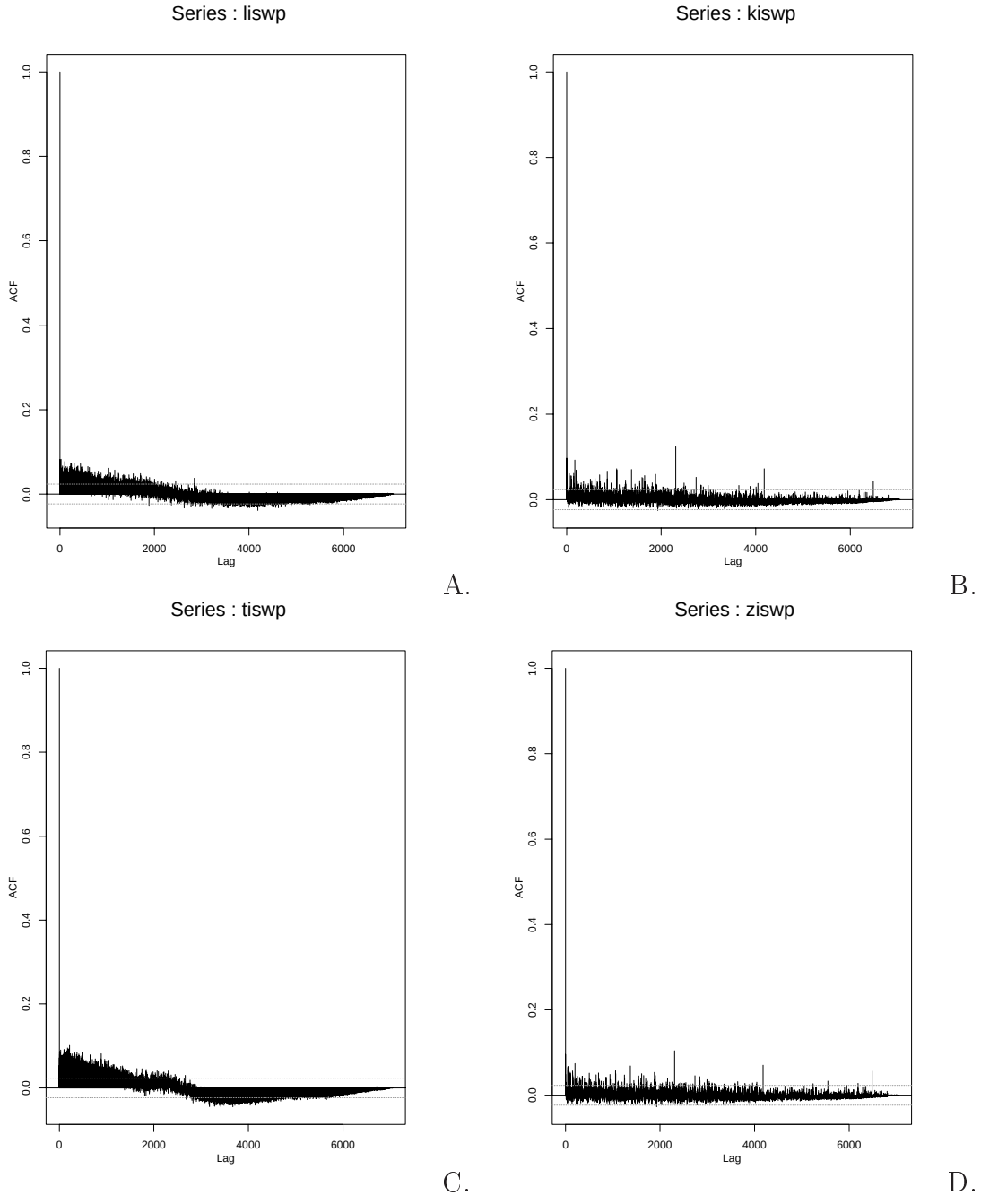


Figure 8: ACF of squared 5m residuals from MP with 200 (A-B) and 500 (C-D) atoms on the WP dictionary for respectively original (left) and de-noised (right) tables.

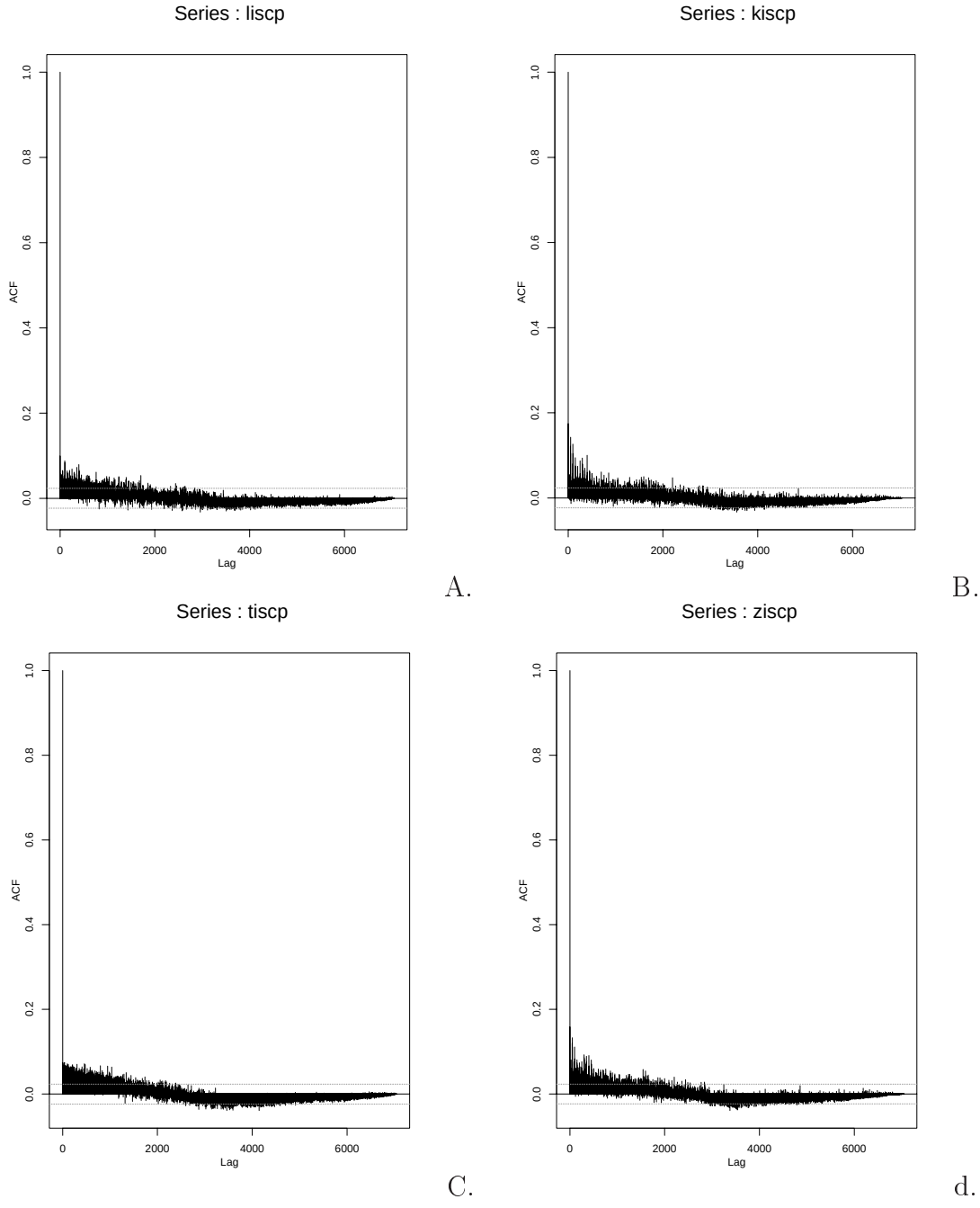


Figure 9: ACF of squared 5m residuals from MP with 200 (A-B) and 500 (C-D) atoms on the CP dicitonary for respectively original (left) and de-noised (right) tables.

T = # of Atoms	50	100	200	500	w50	w100	w200	w500
n=1-3328								
level 0	**0.662**	**0.733**	**0.743**	**0.699**	**0.21**	**0.633**	**0.657**	**0.672**
level 1	**0.065**	**0.055**	**0.06**	**0.119**	**0.268**	**0.024**	**0.017**	**0.176**
level 2	0.005	0.036	0.059	0.078	0.018	0.032	0.09	0.037
level 3	**0.177**	**0.1**	**0.079**	**0.056**	**0.284**	**0.19**	**0.13**	**0.061**
level 4	0.045	0.029	0.022	0.018	0.143	0.032	0.024	0.025
level 5	0.017	0.02	0.018	0.014	0.026	0.014	0.019	0.016
level 6	0.028	0.029	0.02	0.015	0.051	0.075	0.063	0.012
n=3329-7040								
level 0	**0.275**	**0.516**	**0.696**	**0.74**	**0.528**	**0.314**	**0.327**	**0.607**
level 1	**0.581**	**0.353**	**0.226**	**0.185**	**0.079**	**0.151**	**0.178**	**0.112**
level 2	0.033	0.033	0.027	0.028	**0.203**	**0.005**	**0.029**	**0.136**
level 3	0.009	0.013	0.008	0.014	0.092	0.054	0.088	0.081
level 4	0.028	0.025	0.011	0.01	0.072	0.055	0.081	0.034
level 5	0.009	0.019	0.013	0.01	0.012	0.136	0.105	0.015
level 6	0.064	0.041	0.019	0.013	0.015	0.285	0.192	0.005

Table 2: Energy percentage distribution among resolution levels for sub-sampled residual 5m series transformed via CPT and computed via MP algorithm at different degrees of approximation power, i.e. with 50, 100, 200 and 500 atoms. MP runs with the original (left part of the table) and de-noised crystals.

high frequency resolution, somehow regardless the presence of noise, the WP transform seems to require a pre-processing stage of de-noising so to remove the spurious effects of the noise, mostly visible at the finest scales, and thus improving the detection power of the low frequency informative content of the signal in the forms of strong dependencies and periodicities.

5.3 Time and Frequency Resolution Pursuit.

Together with the risk of finding spurious components for the non-stationary nature on the data, the masking effects of noise has been indicated as a further difficulty in dealing with high frequency financial time series. These factors should be considered combined with possible overfitting effects when the MP optimization procedure is run; the algorithm could learn too much and adapt even to non-features. De-noising through thresholding with the SURE-Shrink estimator alone may not be sufficient for optimally dealing with all these aspects. An important aspect concerns the structure of the algorithm itself, and its pursuit activity throughout the resolution levels.

An algorithm known as *High Resolution Pursuit* (HRP) (Jaggi, Karl, Mallat & Willsky, 1998) has been proposed so to improve the local fit power compared to that of MP; the way to do so is by using information just from the highest scales. One can imagine to address each atom of a dictionary through a set of indices, $I_\gamma(k)$ including functions g_γ each formed by averaging elements at finer scale j+k, i.e. $g_\gamma = \sum_{i=1}^{m} \alpha_i g_{j+k,t_i}$[9]. Examples are offered for atoms from B-Spline and WP dictionaries. This new algorithm performs very well compared to MP and *Basis Pursuit* (Chen, Donoho & Saunders, 2001) and presents

[9]A new locally sensitive *similarity measure* has been introduced with the aim of selecting the most informative atoms to be used by the pursuit algorithm, whenever the atoms belonging to low scales can be represented as averages of finer resolution atoms.

clear advantages in some cases, but has limitations as well. It is not fully adaptive with regard to the specific scale selection, thus left to heuristic rules, and to the stopping rule, which is set to avoid overfitting problems, is application or case dependent.
We don't modify or adapt the algorithm itself, but consider HRP as a good premise for understanding how to optimize the selection of information coming from the detail signals obtained by the MRA, with regard to both its time and frequency content. Thus, we investigate the performance of the MP algorithm when is applied on a restricted and ad hoc selected range of resolution levels, i.e. the finest resolution levels of the WP and CP tables, which are obtained through ICA in the WP and in the CP cases. We adopt the same flexible degree of approximation power of 50, 100, 200 and 500 atoms and compare the energy percentage distribution obtained after the MP runs.
In Table 3 below reported we have the two estimated mixing matrices A, where the observed sensor signals are those computed at each resolution levels by the WP and the CP transforms. 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[10].
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 hypotesis that financial markets operate under conditions driven by dynamics which are different according to the time horizons considered for evaluating returns from the invested resources; an example is offered by comparing speculative (short term) and longer term forms of investments, from day-by-day trading to mutual funds or balanced portfolio strategies.

Resol. 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
CP-A							
level 0	0.0029	0.0062	0.008	0.0031	0.0021	**0.1261**	0.0033
level 1	0.0012	0.0033	0.0013	0.0013	0.0041	0.0039	**−0.1204**
level 2	−0.0089	−0.0023	−0.0031	**−0.1712**	0.0031	0.0057	−0.0006
level 3	**0.1868**	−0.0008	−0.0038	−0.0114	−0.0057	−0.0031	0.006
level 4	−0.0022	**0.1832**	0.0011	−0.0002	−0.0191	−0.0083	0.0053
level 5	0.006	0.0142	−0.0059	0.002	**0.1482**	−0.0053	0.0035
level 6	0.0014	0.0046	**0.1748**	−0.0021	0.0036	−0.0052	0.002

Table 3: Weights of the estimated ICA mixing matrix distributed across resolution levels for residual 5m series obtained in WP/CP tables.

Since our sensor signals are obtained from a multi-resolution decomposition of the

[10]This choice is done just for convenience, and not because we want to pre-select their number according to some assumption or a-priori knowledge coming fom the market context.

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. From the CP extimated mixing matrix A things change substantially, since each level depends mainly from out-of-level factors, i.e. the independent components found are not in a diagonal form but belong instead to other resolution levels, remaining only negligibly influenced by within-level factors.
Considering the results obtained with the ICA application, 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 and CP tables, and find a possible explanation or at least some help for how to interpret those findings (see Table 4). The ICA experiment simply works as a test procedure which clearly suggests the goodness of the previous strategy more with the WP table than with the CP table.

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
CP table				
level 0	0.819	0.637	0.704	0.722
level 1	0.021	0.135	0.145	0.150
level 2	0.081	0.127	0.084	0.068
level 3	0.079	0.101	0.067	0.060

Table 4: Energy percentage distribution among the 3 finest resolution levels for residual 5m series obtained in WP/CP tables and computed via the MP algorithm at the approximation power of 50,100,200 and 500 atoms.

We notice that with the WP table level 0 increseas with T and level 3 decreases with T, and they gradually exchange the relative contribution to the total energy, while the other two levels are pretty much similar. For the CP table level 1 increases with T, the other being stable, while level 0 remains the one capturing the biggest percentage of energy. In the next figures we repeat the diagnostic ACF plots already shown before, based on the new residuals; Figure 10 and Figure 11 report the absolute and the squared ACFs for the residuals from the WP and the CP tables.

5.4 Interpreting the Results
In the CP case, levels 4, 5 and 6 mostly depend, respectively, from the specific information content of levels 1, 4 and 2; thus, by including only levels 0-3 in the MP range of application, we reduce the frequency information loss coming from excluding the low scales, and even if still sub-optimally, we obtain a better compromise with regard to the trade-off of time and frequency resolution with which we let MP operate.
As said before, MP benefits because working with least dependent coordinates allow to

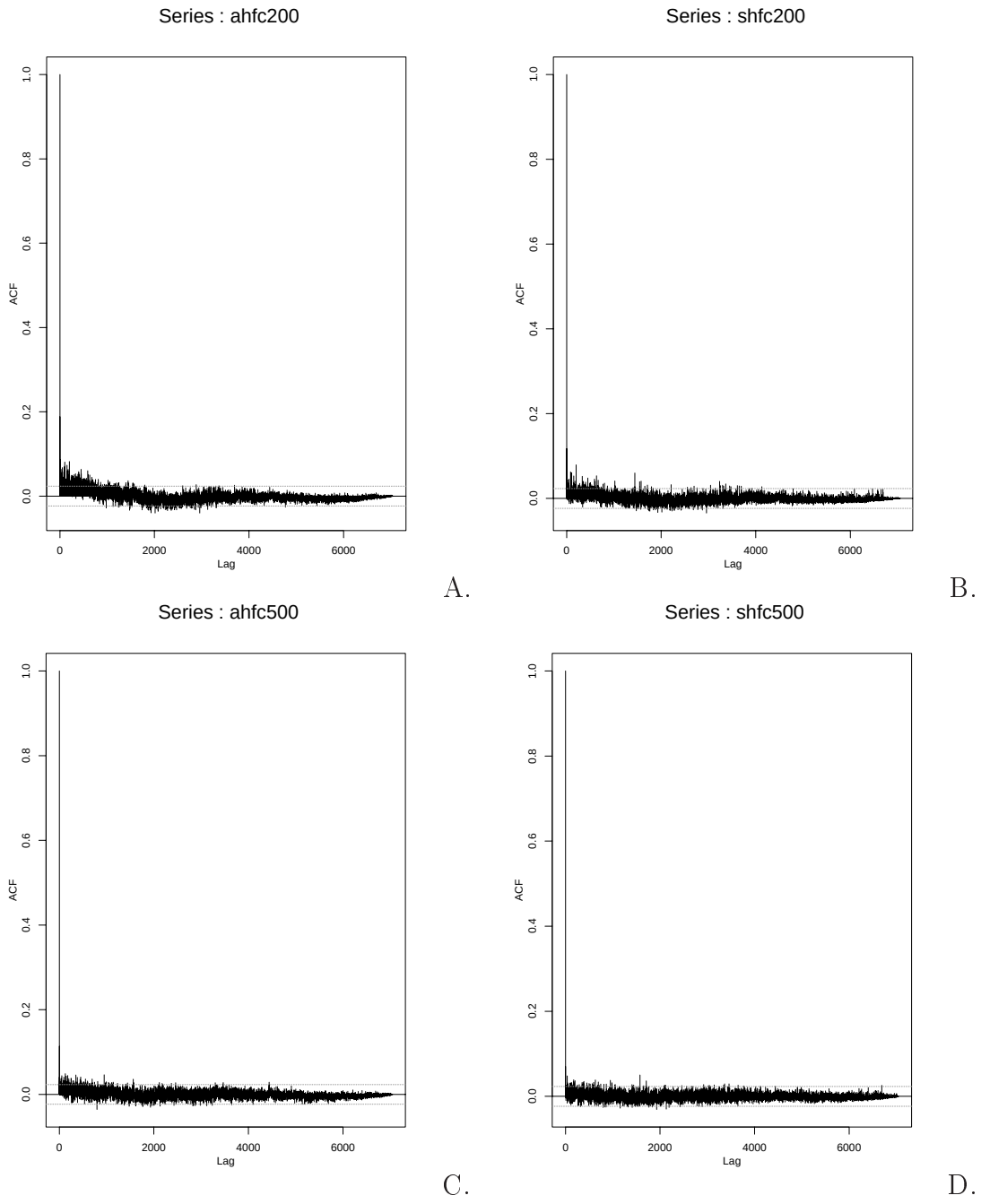


Figure 10: ACF of absolute (A-C) and squared (B-D) 5m residuals from MP with 200 and 500 atoms on the CP dictionary at the finest four resolution levels.

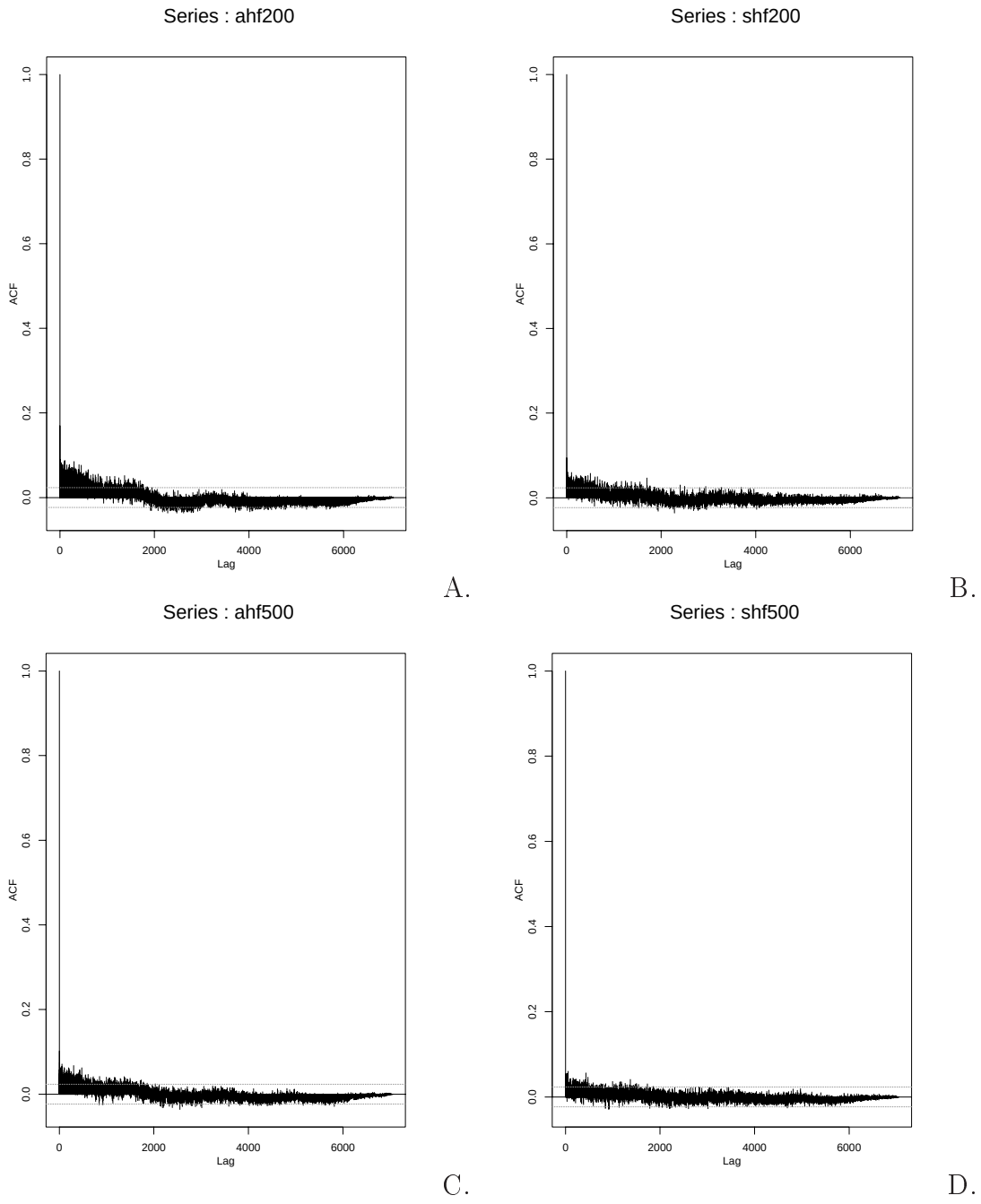


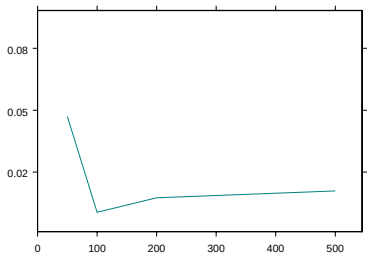
Figure 11: ACF of absolute (A-C) and squared (B-D) 5m residuals from MP with 200 and 500 atoms on the WP dictionary at the finest four resolution levels.

learn faster and better, due to a more orthogonalized algorithm.
We then observe that with the WP table the dependencies left in the ACF plots are less evident than before, particularly with regard to the long memory component, while the initial autocorrelation decreases with T. For the CP table the picture suggests even a better ability of MP to capture and remove these dependencies, thus suggesting that the feature detection power improves qualitatively and with computational savings by simply concentrating the MP activity only on the finest resolution levels.
This fact indirectly addresses the power of the HRP algorithm compared to the MP, when the latter is active on the whole resolutions domain, but our procedure also suggests that one can follow simple strategies instead of modifying the algorithm. In fact, the MP may still be highly successful by just limiting its activity to the finest resolution levels, and particularly in the WP case, by exploiting the information content of high-scale signals compared to the low-scale ones (Jaggi, Karl, Mallat & Willsky, 1998).
The advantages of working with band-pass filtered detail signals in terms of temporal aggregation effects are known (Abry, Veitch & Flandrin, 1998); they are stationarized and decorrelated by wavelets, as seen, in the sense of being almost uncorrelated along individual scales and almost independent across scales. We support our results with other arguments too, as explained below, which explain that the selection of details reflects the selection provided by ICA on the wavelet expansion coefficients, justified on the grounds that the least dependent components lead to more orthogonalized MP and thus better efficiency.
Figure 12 is about the performance of the MP algorithm when examined through the residues obtained at varying approximation power employed. For the case under study, we consider the L_2 and L_1 errors, from respectively squared and absolute transformed residual terms, and compare them with the number of MP approximating, possibly coherent, structures employed, up to 500, which corresponds to the L_0 norm of the expansion coefficients, i.e. a measure of sparsity.

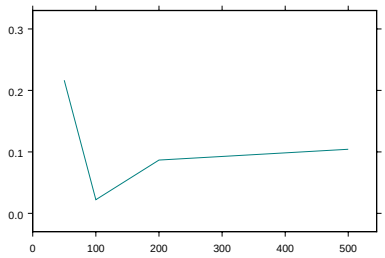
We note that with CP tables the MP has an excellent performance, but in both cases, L_2 and L_1 norms, the first turning point is at approximately 100 structures, while the second one is at approximately 200 structures, and while for the for the L_2 norm is smooth, for the L_1 norm is slightly steeper in the decrease toward the approach to the new minimum at approximately 500. For the WP case instead, the minimum seems reached at approximately 100, and there is no reverting behaviour afterwards, even if with different slopes starting from 200 structures. The plots look very similar for the two norms.
We note that in both cases we don't have a guarantee that if the algorithm is run for more than 500 structures, it will go in one direction or another, due to the risk of overfitting. But while for the CP case we can see that after 200 iterations it stabilizes its pattern and reverts toward the limit reached by 100 structures, and thus we can accept this last number as a good indicator for when to stop for observing the dynamics of the volatility process and control the related unstabilities of MP, for WP instead we should definitely stay with 100 structures so to avoid overfitting and conclude that further iterations would allow for dictionary noise to be encountered in both norms.

5.5 Conclusive Remarks

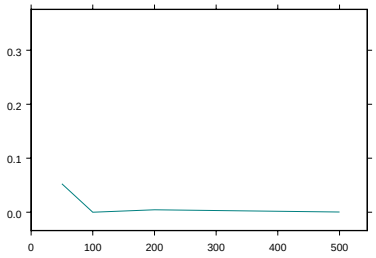
By looking at the results obtained with the high frequency time series application and with the use of WP and CP libraries, various considerations could be advanced. ICA applied to WP based detail signals yields results that best match the search for a com-



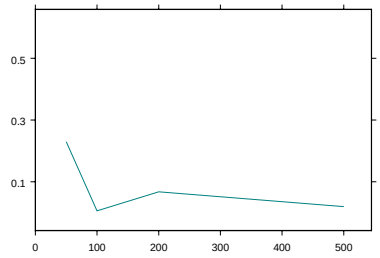
A.



B.



C.



D.

Figure 12: L_2 error vs number of approximating structures, for WP (A) and CP (C); L_1 vs L_0 norm for WP (B) and CP (D).

promise between achieving a sparse representation together with a set of least dependent components.
The selection of high scale signals eliminates redundant information by keeping highly localized time resolution power without simultaneously losing too much frequency resolution, due to the fact that low scale information can be reproduced by averaging high scale one. The denoising step too, here applied, permits to improve the S/N ratio considerably, and thus delivers a sparser signal representation.
The independent components need not to exist, particularly with non-stationary and dependent signals; one must turn to other devices, or combine ICA with wavelet-based signal decomposition and de-noising, so to form a sort of SCA.
For non-Gaussian data one finds that wavelet-represented signals result the least dependent components selected, where the detail sequences are obtained by sequential application of WPT/CPT and ICA. These sequences are sparse, for the choice of ad hoc dictionary selection and for the packet coefficient thresholding stage.
When ICA is applied to a CP library, it doesn't really build a sparse representation, since the CP coordinates are already naturally endowed with that property; from one aspect it depends on the time domain segmentation operated according to the degree on discontinuity revealed by the data. Thus, for a certain time interval, the size of the local cosine windows might correspond well to that representing an approximate stationary behavior for the process at hand.
From (Mallat, Papanicolaou & Zhang, 1998) we know that local cosine vectors might be approximate eigenvectors of the covariance operators and that an orthogonal basis of them yields a sparse matrix with 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 BOB, but we have already seen that for our time series is sub-optimal compared to the greedy MP.
Thus, the least dependent levels and the source separation steps enabled by ICA based on the CP decomposition, now form an hybrid procedure and deliver a mix of components which unlike with WP are not concentrated at the finest resolutions. As far as concerns the independence among resolution levels, there isn't a precise selection order, but instead low and high frequency information content collected at various degree of resolution. In terms of decomposing the signal, the advantage of using a CP transform is thus in the inherent diagonalization power with respect to the covariance operator.
Furthermore, our findings address indirectly the power of HRP compared to MP; however, due to our simple ICA-based procedure of pre-selecting the resolution levels over which MP runs, it also suggests a simple strategy aimed to bypass the use of modified algorithms bringing limitations and contraints into the analysis.
The MP algorithm may be very effective by just limiting its range of activity, in this case the domain of resolution levels obtained from a signal decomposition. Exploiting the independent information content of MRA signals, as indicated by the ICA stage, may represent an efficient procedure and a near-optimal way of tuning the resolution pursuit.

Acknowledgements.

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. Abry, P., Flandrin, P., Taqqu, M.S., & Veitch, D.,. Wavelets for the analysis, estimation and synthesis of scaling data. In Park, C. & Willinger, W. (Eds), *Self-similar network traffic and performance evaluation*, (2000) 39-88, New York: Wiley.
2. Abry, P., Veitch, D., & Flandrin, P.,. Long range dependence: revisiting aggregation with wavelets. *Journal of Time Series Analysis*, 19(3), (1998) 253-266.
3. Amari, S.,. ICA for temporally correlated signals - Learning algorithms. Tech. Report, RIKEN (JP) 1998.
4. Andersen, T., & Bollerslev, T.,. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4 (1997) 115-158.
5. Andersen, T., & Bollerslev, T.,. Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns. *The Journal of Finance* LII(3) (1997) 975-1005.
6. Attias, H.,. Blind Source Separation and Deconvolution: the Dynamic Component Analysis Algorithm. *Neural Computation*, 10 (1998) 1373-1424.
7. Attias, H.,. Independent Factor Analysis. *Neural Computation*, 11(4) (1999) 803-851.
8. Back, A. D., & Weigend, A. S.,. A first application of Independent Component Analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8 (1997) 473-484.
9. Bruce, A. & Gao, H.V.,. *S+Wavelets*, Seattle: StaSci Division, MathSoft Inc. 1994.
10. Capobianco, E.,. Wavelets for High Frequency Financial Time Series. In *Interface '99 Conference Proceedings* (1999) 373-378.
11. Capobianco, E.,. Statistical Analysis of Financial Volatility by Wavelet Shrinkage. *Methodology and Computing in Applied Probability*, I(4) (1999) 423-443.
12. Cardoso, J.,. Source separation using higher order moments. In *Proceedings International Conference on Acoustic, Speech and Signal Processing* (1989) 2109-2112.
13. Cardoso, J., & Souloumiac, A.,. Blind beamforming for non-Gaussian signals.

IEE Proceedings F., 140(6) (1993) 771-774.

14. Chen, S., Donoho D., & Saunders, M.A.,. Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43(1) (2001) 129-159.
15. Coifman, R., & Wickerhauser, V.,. Entropy-based algorithms for best basis selection. *IEEE Transactions in Information Theory*, 38 (1992) 713-718.
16. Comon, P.,. Independent Component Analysis - a new concept?. *Signal Processing*, 36(3) (1994) 287-314.
17. Dahlhaus, R.,. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25 (1997) 1-37.
18. Daubechies, I.,. *Ten Lectures on wavelets*. Philadelphia: SIAM 1992.
19. Davis, G., Mallat, S., & Avellaneda, M.,. Greedy adaptive approximations. *Journal of Constructive Approximation*, 13(1) (1997) 57-98.
20. Davis, G., Mallat, S., & Zhang, Z.,. Adaptive time-frequency approximation with Matching Pursuit. In Chui, C.K., Montefusco, L., & Puccio, L., (Eds), *Wavelets: Theory, Algorithms and Applications*, (1994) 271-293. Academic Press Inc.
21. Donoho, D.,. Unconditional Bases and Bit-Level Compression. *Applied and Computational Harmonic Analysis*, 3 (1996) 388-392.
22. Donoho, D.,. Sparse Components of Images and Optimal Atomic Decompositions. Technical Report 1998-31, Stanford University (US) 2000.
23. Donoho, D., & Johnstone, I.M.,. Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika*, 81 (1994) 425-455.
24. Donoho, D., & Johnstone, I.M.,. Adapting to unknown smoothness via wavelet shrinkage. *Journal of American Statistical Association*, 90 (1995) 1200-1224.
25. Donoho, D., & Johnstone, I.M.,. Minimax Estimation via Wavelet Shrinkage. *The Annals of Statistics*, 26 (1998) 879-921.
26. Donoho, D., Mallat, S., & von Sachs, R.,. Estimating Covariances of Locally Stationary Processes: Consistency of Best Basis Methods. In *Proceedings of IEEE Time Frequence and Time-Scale Symposium*, (1996) 337-340. New York: IEEE.
27. Donoho, D., Mallat, S., & von Sachs, R.,. Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods. Tech. Report 1998-517, Stanford University (US) 1998.
28. Gao, H.Y.,. Wavelet shrinkage estimates for heteroscedastic regression models. Tech. Report, MathSoft Inc. 1997.
29. Girosi, F.,. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6) (1998) 1455-1480.
30. Hardle, W., Kerkyacharian, G., Picard, D., & Tsybakov, A.,. *Wavelets, Approximation, and Statistical Applications*, New York: Springer-Verlag 1998.
31. Hyvarinen, A., & Oja, E.,. A fast fixed-point algorithm for Independent Component Analysis. *Neural Computation*, 9(7) (1997) 1483-1492.
32. Hyvarinen, A.,. Independent Component Analysis for Time dependent Stochastic Processes. In *ICANN'98*, 541-546.
33. Hyvarinen, A.,. Fast and robust fixed-point algorithms for Independent Component

Analysis. *IEEE Transactions on Neural Networks*, 10(3) (1999) 626-634.

34. Jaggi, S., Karl, W.C., Mallat, S., & Willsky, A.S.,. High Resolution Pursuit for Feature Extraction. (1998) *Applied and Computational Harmonic Analysis*, 5(4) (1998) 428-449.
35. Johnstone, I.M.,. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, 9 (1999) 51-83.
36. Johnstone, I.M., & Silverman, B.W.,. Wavelet threshold estimators for data with correlated noise. *Journal of Royal Statistical Society, Series B.*, 59 (1997) 319-351.
37. Jutten. C., & Herault, J.,. Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24 (1991) 1-10.
38. Kisilev, P., Zibulevsky, M., Zeevi, Y., & Pearlmutter, B.A.,. Multiresolution framework for blind source separation. Technical Report, Dept. of Electrical Engineering, Technion, Haifa (Israel) 2000.
39. Krim, H., & Pesquet, J.C.,. On the statistics of Best Bases criteria. In Antoniadis, A., & Oppenheim, G., (Eds), *Wavelets and Statistics* (1995) 193-207, New York: Springer-Verlag.
40. Lewicki, M.S., & Sejnowski, T.J.,. Learning Overcomplete Representations. *Neural Computation*, 12 (2) (2000) 337-365.
41. Li, Y., & Xie, Z.,. The wavelet detection of hidden periodicities in time series. *Statistics and Probability Letters*, 35 (1997) 9-23.
42. Mallat, S., & Zhang, Z.,. Matching Pursuit with time frequency dictionaries. *IEEE Transactions Signal Processing*, 41 (1993) 3397-3415.
43. Mallat, S., Papanicolaou, G., & Zhang, Z.,. Adaptive Covariance Estimation of Locally Stationary Processes. *The Annals of Statistics*, 26(1) (1998) 1-47.
44. Meyer, I.,. *Wavelets: algorithms and applications*. Philadelphia: SIAM 1993.
45. Neumann, M.H., & von Sachs, R.,. Wavelet Thresholding: beyond the Gaussian I.I.D. situation. In Antoniadis, A., & Oppenheim, G., (Eds), *Wavelets and Statistics* (1995) 301-329, New York: Springer-Verlag.
46. Olshausen, B.A., & Field, D.J.,. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37(23) (1997) 3311-3325.
47. Poggio, T., & Girosi, F.,. A sparse representation for function approximation. *Neural Computation*, 10(6) (1998) 1445-1454.
48. Saito, N.,. The Least Statistically Dependent Basis and its applications. In *Proceedings 32nd Asilomar Conference on Signals, Systems and Computers* (1998) 732-736, IEEE.
49. Saito, N., Larson, B.M., & Benichou, B.,. Sparsity and Statistical Independence from a best basis viewpoint. In *Proceedings SPIE, Wavelets Applications in Signal and Image Processing VIII*, 4119, Aldroubi, A., Laine, A.F., & Unser, M.A., (Eds) (2000) 474-486.
50. Serroukh, A., Walden, A.T., & Percival, D,B.,. Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *Journal of the American Statistical Association*, 95(449) (2000) 184-196.

51. von Sachs, R., & MacGibbon, B.,. Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics*, 27 (2000) 475-499.

52. Wu, L., & Moody, J.,. Multi-effect decompositions for financial data modelling. In *NIPS'96*, 9 (1996) 995-1001.

53. Zibulewsky, M., & Pearlmutter, B.A.,. Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neural Computation*, 13(4) (2001) 863-882.

54. Zibulevsky, M. & Zeevi, Y.,. Extraction of a single source from multichannel data using sparse decomposition. Technical Report, Dept. of Electrical Engineering, Technion, Haifa (Israel) 2001.