European Society of Computational Methods in Sciences and Engineering (ESCMSE)

Journal of Numerical Analysis, Industrial and Applied Mathematics (JNAIAM)
vol. 1, no. 2, 2006, pp. 147-174
ISSN 1790–8140

Non-linear Functional Approximation of Heterogeneous Dynamics[1]

E. Capobianco[2]

CRS4 - Polaris,
Science and Technology Park of Sardinia,
Pula (Cagliari), Italy

Received 6 July, 2005; accepted in revised form 17 July, 2006

Abstract: Many applications hinge on modelling phenomena observed or sampled through signal sequences or time realizations whose dynamics come from heterogeneous sources of uncertainty. This requirement applies to many scientific contexts, and results particularly challenging when a low signal-to-noise ratio exists, due to structural or experimental conditions, and the information core appears dispersed in a wide spectrum of frequency bands or resolution levels. The methodological and empirical work that is here presented aims to design ad hoc approximation instruments dealing with a particularly complex class of random processes that generates financial returns, or their aggregates in the form of index returns. It is important to note that the underlying volatility function appears to be subject to the signature of noise and the masking effects of non-stationary superimposed dynamics, together with multi-scaling regimes. Due to the unobservability of the volatility function, its recovery represents an inverse problem that can be cast in a latent variable model designed to account for both switching multi-scaling regime and cascade system dynamics. Together with emphasizing the role of independent component analysis for achieving dimensionality reduction of the addressed inverse problem, we also stress the role of atomic functional dictionaries in improving the volatility feature detection power, and show the performance of greedy approximation algorithms in delivering sparse representations and coherent decompositions of the return sequences.

Keywords: Cascade Systems; Non-linear Sparse and Greedy Approximation; Independent Component Analysis; Multi-scaling; Volatility Recovery

Mathematics Subject Classification: 60H30, 62M10, 62G08, 60G35.

[1]Published electronically October 28, 2006
[2]E-mail: ecapob@crs4.it

1 Introduction

An important part of most scientific methodological work is devoted to the design of model building strategies for representing phenomena whose dynamics depend on heterogeneous sources of uncertainty, variation and perturbation. This heterogeneity applies to many application and experimental contexts, as for instance to return generating stochastic processes where the underlying volatility function may be subject, among other dynamics, to changes in scaling regimes.

The last aspect has been illustrated in recent econophysics and multi-fractal studies, but is not restricted to the financial field. It is often observed from applications in network data traffic, physiological studies of ECG and heart rate data, earthquakes, speech and audio analysis, and many other examples.

The presence of multi-scaling regimes addresses one key fact about the decay rates of covariance and correlation functions of the stochastic processes involved. They usually show power law behavior due to self-similarity and long range dependence. The same functions might be subject to dynamics which vary with the resolution levels through which we observe the realizations. In order to investigate this structural aspect, ad hoc methodological and algorithmic approaches are here suggested.

Despite wavelet studies (Daubechies, 1992; Meyer, 1993) and multi-resolution theory (Mallat, 1989) have had a strong impact in dealing with multi-scaling dynamics, it has appeared that a much less evident emphasis has been given to these instruments in financial studies. Here the data sources are subject to noise and covariance non-stationarity, appear non-Gaussian and ranging from low- to very high-frequency dynamics. This fact seems somehow to fit the inherent nature of financial markets, with a wealth of agents and investors acting through their investment decisions over different time horizons so as to optimize their risk-return strategies and consequently their portfolio performance.

The methodological proposal follows two mainstreams: 1) Exploiting functional series expansion representations of non-linear random signals through cascade system wavelet-based transforms; 2) Employing sparse approximation algorithms so as to emphasize the ability of coefficient sequences to explain the structural features of complex stochastic processes observed under non-standard parametric statistical conditions. The applied part of the paper refers to high-frequency data over which the performance of *ad hoc* computational techniques is measured.

The paper is organized as follows: Section 2 reviews aspects of random processes, while Section 3 introduces sparse approximation techniques, including wavelet transforms, cascade systems and indipendent component analysis. Section 4 illustrates latent variable systems for volatility processes, and Section 5 presents non-orthogonal and overcomplete systems, together with greedy approximation techniques. Section 6 reports experiments about the extraction of volatility source from high frequency data, and shows that multi-scaling regimes can be detected. Section 7 is for the conclusions.

2 Random Processes Review

2.1 General Concepts

The first definition of interest is that of a stochastic process and its realization sequences, respectively addressed by $Y(t)$ and y_t, and with the former possibly extended to $[Y(t,\omega), t \in T, \omega \in \Omega]$, given a probability space $(\Omega, \mathcal{F}, P)$. The (zero-mean, for simplicity) process $Y(t)$ admits a covariance function $COV(s,t) = E[Y(s)Y(s+t)]$ and a corresponding correlation function $COR(s,t) = COV(s,t)/VAR(s)$, where in the denominator it holds that $VAR(s) \equiv COV(s,s)$.

Two well-known facts are: 1) the Kolmogorov's consistency theorem, for which a positive definite form K corresponds to a unique Gaussian process $G_K(t)$ admitting that form as the covariance

kernel; 2) the Moore-Aronszajn theorem, leading to a one-to-one correspondence between K and a reproducing kernel Hilbert space (RKHS) R_K with reproducing kernel K. The RKHS, which is a Hilbert space in which all the point evaluations are bounded linear functionals, or in other terms spaces of functions for which point-wise values are defined, results very relevant for the class of projection operators that are adopted in this paper.

2.2 Decompositions

The simplest process decomposition is *a la* Parzen (1961, 1970), which under both stationary and non-stationary moment conditions merely requires a representation for the observable realization as mean value (m) and fluctuation (f) functions, i.e., $y_t = m_t + f_t$.

We aim to investigate stochastic processes that instead of being globally stationary are just locally stationary, or even non-stationary. Dalhaus (1997) first, then Nason *et al.* (1999; 2000) have proposed formal approaches to define locally stationary processes. The basic idea of local stationarity is that random processes with this statistical property can always be characterized, simply speaking, by an approximately stationary behaviour over some observed time intervals.

A statistical model is often built under non-standard conditions; this occurs with processes typically characterized by non-stationarity in the first two conditional moments of the distributions of interest, or by non-gaussianity, heteroskedasticity, self-similarity and even other conditions that complicate inference for the underlying stochastic dynamics, and usually lead to more or less severe model mis-specification.

Some insight is offered by the eigen-system representation of stochastic processes, according to which one may exactly or approximately adapt to, respectively, stationary or non-stationary statistical conditions.

By the Mercer's theorem (Grenander, 1981), there exist orthonormal sequences of eigenfunctions e_i and eigenvalues λ_i, such that $\lambda_1 \geq \lambda_2 \geq \ldots 0$, and the following kernel decomposition holds with absolute and uniform convergence, $R(.,.) = \sum_i \lambda_i e_i(.)e_i(.)$. For a second-order process continuous in the mean this orthogonal basis representations is associated to the covariance function and leads to the Karhunen-Loéve basis expansion for which convergence occurs in the mean in $L_2(.)$, and such that the following decomposition applies:

$$Y(t) = \sum_i y_i e_i(t) \tag{1}$$

with

$$y_i =< Y(t), e_i(t) > \tag{2}$$

or equivalently $y_i = \int Y(t)e_i(t)dt$.

In terms of decomposition power, we range from the simplest case of Gaussian processes for which the second order dependence structure is known and delivers the most efficient basis through a diagonalization of the autocovariance matrix, to the case where the latter matrix is unknown, and the class of processes under study is non-stationary and can only be approximated by eigen-system solutions.

Thus, the expected optimal covariance diagonalization that can be achieved with the Fourier basis tranform under suitable conditions, can also be just sub-optimally obtained under different conditions and by using other bases.

3 Sparse Approximation Techniques

3.1 Wavelet Transforms

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 such that the terms indicated with ψ_{jk} generated by $j-$dilations and $k-$translations form orthonormal bases through $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$, one may compute differences among approximations related to successively coarser resolution levels.

For $f \in L^2(R)$ (with $< ., . >$ the L^2 inner product) and given an analyzing (admissible) wavelet with its doubly-indexed generated family, the continuous wavelet transform WT^c is (Daubechies, 1992):

$$WT^c(f)_{jk} =< f, \psi_{jk} >=\mid j \mid^{-\frac{1}{2}} \int f(t)\psi(\frac{t-k}{j})dt \tag{3}$$

Given the constant c_ψ, the function f can be recovered from the reconstruction formula:

$$f = c_\psi^{-1} \int \int WT^c(f)_{jk}\psi_{jk}\frac{djdk}{j^2} \tag{4}$$

Corrispondingly, a sequence of smoothed signals and of details giving information at finer resolution levels is found from the discrete time wavelet transform WT^d of the process realization. This leads to the following decomposition:

$$f = \sum_l c_{m_0 l}\phi_{m_0 l} + \sum_{m>m_0}\sum_l d_{ml}\psi_{ml} \tag{5}$$

where $\phi_{m_0 l}$ is associated with the related coarse resolution coefficients $c_{m_0 l}$, and d_{ml} are the detail coefficients[3]. In short, the first term of the right hand side is the projection of f onto the coarse approximating space V_{j_0}, while the second term represents the cumulated details.

3.2 Reproducing Kernel Property

In general, a RKHS is the space of solutions $\upsilon(b, a)$, with b translation and a scale parameters, of an integral equation like:

$$\upsilon(b', a') = P_\psi \upsilon(b', a') = \int_{-\infty}^{\infty} K_\psi(b', a', b, a)\upsilon(b, a)\frac{da}{a^2}db \tag{6}$$

The reproducing kernel is given by:

$$K_\psi(b', a', b, a) = \frac{1}{c_\psi} < \psi_{ba}, \psi_{b'a'} > \tag{7}$$

For WT^d operators, the RKHS formulation involves the idea of implementing a direct and inverse transforms sequentially, from respectively the function pre-image and image spaces (Mallat and Zhong, 1992).

Thus, the redundancy appears through the reproducing kernel by the composition of a direct and of an inverse operator $P_V = W \circ W^{-1}$, with V the space of the WT^d W of the function $f \in L^2(R)$, and P_V a convolution operator for the inverse and the direct WT^d. When $F(j, k) \in$

[3] In general, coarse and detail resolution coefficients are derived from the corresponding wavelet functions, $c_{jk} = \int f\phi$ and $d_{jk} = \int f\psi$

$L^2(Z \times R)$, meaning that a dyadic wavelet transform works through a scale variable sampled on a dyadic grid, the following holds (Carmona *et al.*, 1998) so as to describe the action of the reproducing kernel on F:

$$K_\psi^F(j,k) = \sum_{j'} \int < \tilde{\psi}_{k'}^{2^{j'}}, \psi_k^{2^j} > F_{j'k'} dk' \tag{8}$$

with $\tilde{\psi} \neq \psi$ a dual wavelet defined in the Fourier domain and with wavelet coefficients computed as $< f, \psi_k^{2^j} >$, for any $f \in L^2(R)$. The K_ψ action comes from the inner product of the wavelet and the dual wavelet transforms; the latter defines an inverse transform and is thus used in the expansion formula:

$$f(.) = \sum_j \int < f, \psi_k^{2^j} > \tilde{\psi}_k^{2^j}(.) dk \tag{9}$$

For a WT^c (whose sampling outcome is WT^d) a reproducing kernel operator K_ψ embeds the correlation structure here expressed respectively in theoretical and empirical terms:

$$E[WT^c(f)_{jk}(WT^{c*}(f)_{j'k'})] = c_\psi K_\psi(j,j',k,k') \tag{10}$$

$$K_\psi \approx < \psi_{j'k'}, \psi_{jk} > \tag{11}$$

Following Daubechies (1992) the redundancy of the wavelet transform is reflected by the reproducing kernel property to the image of the original space obtained through the transform. Thus, the initial integral equation (4) can lead through the jk-wavelet transform mapping $f \rightarrow F$ to the space:

$$H_f = \{F \in L^2(R^2), P_f F = F\} \tag{12}$$

where P_f is the integral operator[4] with reproducing kernel K_f given as in (5) by:

$$K_f(j,k,j',k') = \frac{1}{c_\psi} < f_{(j'k')}, f_{(jk)} > \tag{13}$$

Due to the correlation structure accounted by the reproducing kernel K_ψ, wavelets bases act almost as eigenfunctions and only an approximate diagonalization of the covariance structure can be achieved.

As a more general result, the correlation core of the data can be tracked while it changes through a transform a dot-product from $< x_i, x_j >$ to $< \rho(x_i), \rho(x_j) >$, where the latter involves a general non-linear mapping from the input to the feature (image) space. This aspect is further investigated in our experiments.

3.3 Non-linear Expansions

One goal of this paper is to model the mean value of the observed sequences by a small set of basis functions, which means in a sparse way. This criterion satisfies certain requirements in the spirit of many recent proposals; sparse approximations based on special families of basis functions offer feasible solutions in Donoho *et al.* (1998) and in Mallat *et al.* (1998).

[4]Equivalently, it could be considered the orthogonal projection onto H_f, i.e., $P_f^* = P_f^2 = P_f$.

Non-linear process specifications can work through computationally simple models; one example is the well-known signal plus noise framework, also seen as a non-parametric regression setting. The design of optimal algorithms is also strictly dependent on the adoption of adaptive signal approximation techniques, built on the sparsity principle.

By sparse approximation of a certain stochastic process we mean building a model with relatively few parameters embedding the information related to the main features of the underlying process dynamics. Series expansions in some bases or redundant function families are generally used for this scope; depending on the characteristics of the chosen family and on the approximation algorithms employed, one gets interpretable models, computationally efficient, and numerically accurate performances.

Consider a process decomposable in Volterra series:

$$\begin{aligned} Y(t) &= \sum_{\tau_1=-\infty}^{\infty} h_1(\tau_1)X(t-\tau_1) + \\ &\sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} h_2(\tau_1,\tau_2)X(t-\tau_1)X(t-\tau_2) + \ldots \end{aligned} \tag{14}$$

where the functions $h(.)$ are time invariant Volterra kernels. Even if they characterize linear, quadratic and higher order interactions, closed-form expressions hold only for Gaussian inputs and there are high computational requirements due to the infinitely many parameters. It is thus hard to get simple and general identification procedure.

A first example of sparse approximation in relation to a Volterra representation is to consider an Hammerstein series expansion; it improves over the Volterra's, both in mathematical tractability and in computational cost:

$$Y(t) = \sum_{\tau=-\infty}^{\infty} g_1(\tau)X(t-\tau) + \sum_{\tau=-\infty}^{\infty} g_2(\tau)X(t-\tau)^2 + \sum_{\tau=-\infty}^{\infty} g_3(\tau)X(t-\tau)^3 + \ldots \tag{15}$$

with $g(.)$ as the time invariant Hammerstein kernels which model the non-linear system's dynamics by separating each memory term. The consequence of this transform is that the problem of system identification now simplifies, at least in part, since it remains to deal with the block-oriented system structure of the series expansion.

3.4 Cascade Systems Processing

Our next step is to adapt the series representation in (15) so as to maintain a good approximation power and obtain at the same time a sparse model representation of the system's dynamics. From the Hammerstein series structure, a time-varying Hammerstein model can be proposed, following Ralston *et al.*, (1997) and Hasiewicz (2001). A compact structure is given according to:

$$Y(t) = \Lambda\{G[X(t)]\} = \sum_{\tau=-\infty}^{\infty} G[X(t-\tau), t, \tau] \tag{16}$$

which agrees with the general diagram:

$$Input[X] \rightarrow \mathbf{Non-Linear}\ \ system[G] \rightarrow \mathbf{Linear}\ \ system[\Lambda] \rightarrow Output[Y]$$

It is important to consider the inverted expression of the above Hammerstein system; when the input signals are unknown, the recovery procedure starts from the observed output and proceeds backward along the system block-oriented dynamics. One thus gets a Wiener system such as:

$$Z(t) = F\{\Gamma[S(t)]\} \tag{17}$$

which works according to the general scheme:

$$Input[S] \rightarrow \textbf{Linear} \;\; system[\Gamma] \rightarrow \textbf{Non} - \textbf{Linear} \;\; system[F] \rightarrow Output[Z]$$

When the Hammerstein series is expanded with the kernels expressed in a polynomial basis sequence, a correspondent signal+noise version is:

$$Y(t) = \sum_{n=1}^{N} \sum_{m=-M}^{M} g_n(t,m) X(t-m)^n + \epsilon(t) \tag{18}$$

With Hammerstein kernels some linear combinations of basis functions are adopted; examples might be wavelets, harmonics as in Fourier analysis, splines, frames or other approximating structures, say $\psi_k(t)$. In dealing with non-stationary series it is wise to use compactly supported functions, due to their good localization power and the benefits in achieving adaptivity to inhomogeneous smoothness. Accordingly, the revised model becomes:

$$Y(t) = \sum_{n=1}^{N} \sum_{m=-M}^{M} \sum_{k=0}^{K} \beta_n(k,m) \psi_k(t) X(t-m)^n + \xi(t) \tag{19}$$

and its estimation can be conducted under no particular conditions, say without assuming Gaussian density functions and/or stationarity of input processes. The analogous integral functional form for a continuous time version of the model is (Greblicki, 2000):

$$Y(t) = \int_{-\infty}^{\infty} \Upsilon(t-\tau) K[X(\tau)] d\tau \tag{20}$$

which can then be discretized in a compact form as follows:

$$Y(t) \approx \sum_{i=1}^{N} \sum_{\gamma=0}^{l} \Upsilon_i(t,\gamma) K_i[X(t,\gamma)] + \epsilon(t) \tag{21}$$

using an approximation of time-varying dynamics, truncated memory and finite-support functions $K[.]$.

This system depends on the specification of the Υ dynamics and of $pdf(\epsilon)$, eg. the unknown error probability density function. Up to now, the statistical model that one can use for statistical inference purposes requires no restrictions and thus can be parametric, semi- or non-parametric, depending on *a priori* knowledge and on model assumptions.

The random process $X(t)$, input of the cascade system, may just have a bounded variance despite an unknown density function, and still be subject to non-parametric statistical estimation. The non-linear subsystem $K[X(t)]$ is usually unobservable, and assuming $K(.)$ bounded and invertible inference could be conducted. The linear component $\Upsilon(.)$ brings some indeterminacy in the system, when the non-linear characteristics can be estimated only up to a scalar, but the same linear structure can be estimated, as suggested in our approach.

The model identification conditions which refer to possible L^2 multi-scale decomposition and estimation, usually require that high density input regions are selected, but for estimation of inhomogeneous smoothness function classes other aspects become relevant and new instruments have to be used, for instance thresholding (see the papers by Donoho and Johnstone, 1994-1998).

The conditions lead to a functional regression problem, due to the independence of signal and error in the model, such that $E[Y(t+\tau) \mid X(t) = x]$. Standard parametric models are not applicable if the scope is to maintain a flexibility in the model toward properties like localization, sparsity and resolution ability. Kernel estimation and orthogonal series principles find an optimal bridge in wavelets, whose fast convergence applies for large class of input processes and adaptively for non-linearities through the thresholding devices (DeVore, 1998).

3.5 Independent Component Analysis

As said before, recovering the non-linear structure is the hardest part, but it becomes feasible in the Hammerstein setting through the regression set-up. However, this possible recovery may be achieved just up to some accuracy which depends on the linear sub-system estimation.

One possible way to estimate the linear component of a cascade dynamics system is to adopt blind deconvolution techniques like independent component analysis (ICA), from Cardoso (1989) and Comon (1994), and sparse component analysis, from Lewicki and Sejnowski (2000), Donoho (2001), Zibulewsky and Pearlmutter (2001). Ideally, one attempts to combine the advantages delivered by sparsity and independence of signal decomposition, which transfer to better model estimation, combined with signal compression and reconstruction power.

The goal of ICA is 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 through the structure of the index series. The combination of this goal with that of searching for sparse signal representations suggests hybrid forms of sparse component analysis.

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 $n \times m$ mixing matrix A, where usually $m << n$, with n the number of sensor signals, but with $m = n$ holding in many cases too.

The separating or de-mixing matrix $B = A^{-1}$ allows to obtain the $Y = BX$ values, and under a perfect separation $Y = BAS = S$. In real cases the solution holds only approximately, thus solutions are achieved up to permutation P and scaling D matrices, such that $Y = DPS$. The de-correlation and rotation steps which have to be implemented will deliver a set of approximate m independent components.

Independent components can be efficiently computed by ad-hoc algorithms such as joint approximate diagonalization of eigenmatrices for real signals (*JadeR*) (Cardoso and Souloumiac, 1993). 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.

4 Latent Variable Systems for Volatility Processes

4.1 Volatility Definitions

Given a realization sequence r_t, with $t \in (0, T)$ of a return generating stochastic process $\mathcal{R}(t)$, the underlying volatility processes describe second order dynamics of realizations of martingale differences:

$$\epsilon_t = r_t - E[r_t \mid \mathcal{F}_{t-1}] \tag{22}$$

where the increasing filtration of σ-fields stands for the memory of the process, from the past to time $t-1$. Thus, let $\mathcal{F}_{t-1}$ be the information available through the observed data. Linear and non-linear model formulations have been proposed, and parametric, semi-parametric and non-parametric statistical inference have been pursued in volatility studies (see Engle, 1982; Bollerslev, 1986; Ghysels *et al.*, 1996).

Given the usual assumption $\epsilon_t/\mathcal{F}_{t-1} \sim N(0, h_t)$, the conditional variance process h_t could be not stochastic but simply time-varying, according to the past information set $\mathcal{F}_{t-1}$, or could be stochastic, and thus independently driven by disturbance processes.

Due to the fact that in empirical work daily squared returns do not help too much in forecasting the latent volatility structure, the concept of realized volatility, simultaneously studied by Andersen and Bollerslev (1998) and Barndorff-Nielsen and Shephard (2001), has been suggested so to obtain a more accurate estimate of the function with high frequency observations.

The realised volatility is $\hat{\sigma}_t^2 = \sum_{i=1}^{T} r_{i,t}^2$, and thus computes an approximation to the integral (over a certain fixed interval) of the unobservable variable by averaging a certain number of intraday values $r_{i,t}^2$. This leads to $\hat{\sigma}_t^2 \rightarrow \sigma_t^2 = \int \sigma^2(s)ds$, meaning that the integrated volatility is approximated by the realised volatility.

It holds, due to the quadratic variation principle, a result like the following (Karatzas and Shreve, 1988): given X and the partition $T = \{t_0, t_1, \ldots, t_n\}$ of $[0, t]$, the p^{th} variation of X over T is $V_t^{(p)}(T) = \sum_{k=1}^{n} \mid X_{t_k} - X_{t_{k-1}} \mid^p$. Then, if $\mid\mid T \mid\mid = max_{1 \leq k \leq n} \mid t_k - t_{k-1} \mid$ goes to 0, for $p = 2$, the case of interest here with volatility, we have $lim_{||T|| \rightarrow 0} V_t^{(2)}(T) =< X >_t$, where the limit is the quadratic variation of X.

In turn, convergence in probability holds, i.e., $\sum_{j=1}^{n} r_{n,j,t}^2 \rightarrow_{n \rightarrow \infty} \int_0^1 \sigma^2(t+\tau)d\tau$, where the cumulative squared high frequency returns are employed rather than the daily values, so to improve the volatility prediction power.

4.2 Multi-resolution Volatility Decomposition

One faces two problems when approximating the realised volatility and estimating the model parameters involved: the role of smoothness and the presence of noise. Following Antoniadis *et al.* (1995), we might rely on quadratic information from the data, leading to non-negative estimators of the following kind:

$$\tilde{\sigma}_t^2 = \sum_i \alpha_i r_{i,t}^2 \tag{23}$$

for $\sum_i \alpha_i = 0$ and $\sum_i \alpha_i^2 = 1$. This can just be an initial estimate for a more calibrated and robust procedure, since it can be improved by estimators that better account for smoothness and sparsity. Consider the L^2 wavelet decomposition for the volatility function, expressed at time t through inner products:

$$\sigma_W^2 = \sum_k < \sigma^2, \phi_{j_0,k} > \phi_{j_0,k} + \sum_{j > j_0} \sum_k < \sigma^2, \psi_{j,k} > \psi_{j,k} \tag{24}$$

where a smooth part centered on the scaling function is combined with a cumulated sequence of details obtained at different scales from the mother wavelet family. We must then apply the same decomposition to $\tilde{\sigma}^2$, being σ^2 unobservable, and can obtain a perturbed version of the previous approximation:

$$\hat{\sigma}^2 = \tilde{\sigma}^2_W + \epsilon \tag{25}$$

where $\epsilon = W\xi$ is a transformed disturbance, preserving the characteristics of the noise ξ affecting the squared returns. In order to sparsify the estimator, we can apply noise shrinkage techniques or use basis function expansions or overcomplete dictionaries.

Volatility processes have observable and unobservable structure, and thus require quite complex functional class specification to deal with volatility curve and structure recovery. Conversely, properties such as inhomogeneous function classes characterization, covariance diagonalization, de-noising, sparsity and multi-resolution power yield a sound justification for selecting wavelets as approximation and estimation techniques in volatility studies.

If instead of considering the L^2 elements we consider other function classes more suitable for inhomogeneous behavior we can build non-linear estimators for the volatility function through wavelet-type families that inherently account for localization and sparsity. Recent results on volatility applications proposed by Capobianco (1999; 2002a,b; 2003a,b) suggest that a multiresolution view like that allowed by wavelets and related families can be very informative and useful for statistical inference purposes.

4.3 Within- and Across-scale Dynamics

Following Arneodo *et al.* (1998) and the widely documented analogies between price dynamics and hydrodynamic turbulence studies, multiplicative cascade models can be suggested for modeling financial return dynamics. Thus, $r^s_t = \sigma^s_t \epsilon_t$, with $\sigma^s_t = \sqrt{h^s_t}$ and $s < \bar{s}$, the bound being the integral scale. This system is combined with a volatility multiplicatively decomposed over the sequence of scales s_j, for $j = 0, \ldots, J$, $s_0 = \bar{s}$ and s_J the finest resolution level, thus leading to the following cascade representation:

$$\sigma^s_t = \Pi^{J-1}_{j=0} \lambda_{s_j} \sigma^{\bar{s}}_t \tag{26}$$

where the λ_{s_j} random factors link each scale. Thus, it is implied by the model that volatilities at different scales are measured proportionally to each other according to some probabilistic law.

The topic of random cascades is not pursued in the present work, but from the model building perspective adopted the focus goes on the expansion coefficient sequences which are sparsely represented and decomposed in separated sources through ICA. Therefore, the scaling properties of volatility are explored when more independence is brought within the coordinate system.

In other words, the original returns are subject to a new decomposition which leads from the global latent volatility structure to localised components which are transformed and scaled, because embedded in projected detail signals. Therefore, one goal is to show how the compression and the decorrelation power of the wavelet transforms can be supported by a decomposition in independent or least dependent coordinates obtained with ICA.

4.4 Model Specification

The analysis starts by presenting a generalized specification of volatility processes through latent variables. Then, the interest will be in discovering the possible existence of volatility regimes, i.e., regimes characterized by different scaling structure of the volatility process.

The observed return system dynamics are modelled as follows (at time t):

$$r = AS_h + \xi \tag{27}$$

$$S_h = C\Phi + \eta \tag{28}$$

where A, C, and Φ are all time-dependent and r are the observed financial returns computed as $r_t = ln[p_t/p_{t-1}] \times 100$, with p_t the available prices. In particular, A is an unknown linear mixing matrix considered an approximation to a non-linear underlying relation between observed (r) and latent (S_h) variables.

While $\xi \sim i.i.d.(0, \sigma_\xi)$ is a noise process, with σ_ξ representing volatility, S_h are unknown risk factors which independently drive the volatility process and are decomposed in atomic structures of a function dictionary Φ. Then, there are also independent dynamics represented by η.

In general, $h = \sigma_\xi^2 = f[S_h] \approx f[C\Phi]$, for f unknown, can be considered a ridge-like approximation form of the volatility process underlying the returns dynamics. This structure could be further reduced to a weighted linear combination of risk factors, say $h \approx \sum_i w_i \mid S_{h,i}^{risk} \mid^\gamma + \epsilon$, with $\gamma > 0$ and ϵ the approximation error.

The system can embed many volatility specifications, especially after some adjustments on the building block structure of the Φ dictionary; compared to volatility measures of global risk, it puts more emphasis on local measures related to individual risk sources.

Thus, S_h are taken as sources of volatility, specified in a completely non-parametric way and such that a latent volatility process can be still specified following one of its realization[5] and conditionally on the observed return series, or otherwise according to a data generating mechanism subject to more complex stochastic dynamics.

A mixture of random sources behind the volatility dynamics is thus the main structure underlying the presented system; this model strategy is indeed quite classical (Clark, 1973). The novelty comes from the fact that the system dynamics in our model are subject to a mixing mechanism that holds at each scale.

The volatility process can thus be expressed in ridge approximation form. By looking at look at equation (26) and fixing the scale s, say λ_s, we have $\sigma_t^s = \lambda_s \sigma_t^{\bar{s}}$; thus, when combining this feature with a within-scale mixture structure, across-scale cascade dynamics can be set by varying s.

Another interesting aspect of this model is that the sources S_h may have sparse decomposition through Φ. The function dictionary can deliver either a basis or an overcomplete representation[6] for the signal under investigation. The expansion coefficients are indicated by C, while η is an $i.i.d$ process, with no constraints on the probability distribution.

4.5 Methodological Analysis

Since the sources of volatility are unobservable, the goal is to estimate them, together with the mixing matrix. These are quite complicated tasks. One can either build an optimization system with a regularized objective function through some smoothness priors, so to estimate the parameters involved, or can proceed recursively, in the mean square sense, through some iterations of a greedy approximation algorithm.

A special case is when a dual system can be formed and a basis is obtained. In that case the system changes according to the transform $\Phi^{-1} = \Psi$ and as a direct consequence one finds at each time t:

$$S_h\Psi = C\Phi\Psi + \eta\Psi \Rightarrow \tilde{S}_h = C + \tilde{\eta} \tag{29}$$

[5] An implied assumption is thus that of ergodicity, i.e., the fact that space/time averages converge to their ensemble counterparts with an increasing sample size.

[6] See Lewicki and Sejnowski (2000) and Chen *et al.* (2001) for seminal work.

It follows that:

$$Y = AC + A\tilde{\eta} + \xi \Rightarrow Y = AC + \zeta \tag{30}$$

with $\zeta = A\tilde{\eta} + \xi \equiv A\eta\Psi + \xi$. A new system is thus found:

$$Y = AC + \zeta \tag{31}$$

$$\tilde{S}_h = C + \tilde{\eta} \tag{32}$$

If the *signal-to-noise* (S/N) ratio is high with regard to the stochastic nature of the sources of volatility, then $\eta \approx 0$ and $\zeta = \xi$, implying that the same volatility process initially described is found.

If instead S/N is low, the square root of the volatility process becomes characterized by $\Sigma = D + \sigma_\xi$, where $D = A\sigma_\eta\Psi$.

As a result, from a statistical perspective, the volatility structure is expressed non-parametrically through the previous system, with the mixing A now acting on the expansion coefficients C. From the signal (function) space one can now look at the sequence (feature) space, switching from the continuous and/or discrete time functions characterizing the former to the sequences of expansion coefficients typical of the latter.

The idea pursued in the present work is processing the observed returns with an n-greedy approximation and de-noising technique. We build a link between a signal + noise version of the Hammerstein series representation and the atomic dictionaries of approximating structures consisting of wavelet packet (WP) and cosine packet (CP) libraries (see, respectively, Coifman *et al.*, 1992; Meyer, 1993). The following general approximation scheme will be used:

$$Y \approx P\Phi + \xi = AC\Phi + \eta \tag{33}$$

where the noise term η is now including an approximation error from the system together with the measurement effects.

Even if a greedy algorithm already works by sparsifying P through overcomplete representations, and thus de-noising is obtained, it cannot separate the components of the operator P. Another decomposition step is required and in fact we show how well ICA can deal with this aspect. While the A mixing matrix accounts for modulating the dependence structure of the latent volatility sources, the WP and CP expansion coefficients represent the inputs for the ICA step in feature space.

Thus, recovering the volatility function is here seen as an inverse problem which can find a characterization according to Hammerstein-Wiener system dynamics and a solution by using blind signal deconvolution techniques so summarized:

- volatility dynamics are described via convolutive/mixed signals;
- the system representation follows typical Hammerstein-Wiener cascades;
- the non-linear dynamics may benefit from multiscale approximation;
- the procedure designed for an indirect recovery of the volatility structure works through a framework consisting of three main building blocks: atomic (WP and CP) dictionaries, greedy approximation, and space dimensionality reduction via ICA;

- it is useful to explore the estimated $\hat{\eta}$ from (33) for volatility modeling and autocorrelation diagnostics.

According to a generalized Hammerstein-Wiener system schemes, where compared to (16-17) some noise is allowed to enter the system dynamics (non-linear and linear) such that unobservable system components, say UC, are now accounted for, we have the following representation (where NL and L mean non-linear and linear functions, respectively) linking together ICA and greedy algorithm (33):

1. Wiener System: $UC = NL\{L[Input]\}$, with $Output = UC + noise$;
2. Hammerstein System: $Y = L^{-1}\{NL^{-1}[Output]\}$;
3. replace UC for the $Output$ of the previous formula and obtain $Y = Input$, due to the invertibility conditions on the subsystems and the Hammerstein-Wiener inverse relation;
4. the latent component is thus approximated by the greedy algorithm in (33), according to $UC \approx AC\phi$

5 From Orthogonal to Non-Orthogonal Systems

5.1 Other Wavelet Families

The WT^d transform previously introduced is simply a map $f \rightarrow w = Wf$ from the signal domain to the projected domain of wavelet coefficients. While there are clear advantages from an orthogonal wavelet expansion, when combined with usual standard Gaussian assumptions underlying the system, other families of non-orthogonal functions have been suggested and employed. Two main examples are the biorthogonal wavelets and the undecimated wavelet transforms WT^u.

In the former case dual sets of functions, resulting non-orthogonal within each set but orthogonal between the sets, are the building blocks of the approximation systems that find successfull examples of application in inverse problems (see Johnstone, 1999).

With the WT^u transform the matrix W is this time $\bar{N} \times N$, $\bar{N} \geq N$, for which there exists a pseudo-inverse transform matrix W^- such that $W^-W = I$. For a discretized system $d = f + z$, WT^u delivers the following decomposition:

$$Wd = Wf + Wz \rightarrow \tilde{d} = \tilde{f} + \epsilon \tag{34}$$

While the Gaussian property is preserved, i.e., $\epsilon \sim N(0, \Sigma)$, with $\Sigma = WW'$, despite the presence this time of structure in the covariance, i.e., $\Sigma = WW' \neq I$, and of heteroscedasticity when a time-dependent scale factor is considered in the covariance structure.

A key fact is that in the wavelet coefficients domain, a Gaussian stationary and short dependent (or weakly autocorrelated) process is decomposed in a series of detail signals which are independent across scales and temporally uncorrelated along each resolution level (Johnstone and Silverman, 1997). The decorrelation power of wavelets is partially diminished when we deal with non-Gaussian, self-similar and long range dependence processes. However, other systems of wavelet transforms can accomplish the task of decorrelating and stationarizing the series.

5.2 Overcomplete Representations

Function dictionaries are collections of parameterized waveforms (see Chen *et al.*, 2001) that are available from many classes of functions, or 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 WP structures, which represent an extension of the wavelet transform to a richer class of building block functions that allow for a better adaptation due to an oscillation index *osc* related to a periodic behaviour in the series and deliver a richer combination of functions that simultaneously localize a process in time, frequency and scale domains.

For compactly supported wave-like functions $W_{osc}(t)$, finite impulse response filters of a certain length L can be used, and by P-partitioning in (j, osc)-dependent intervals $I_{j,osc}$ one finds through $\{2^{-\frac{j}{2}} W_{osc}(2^{-j}t - k), k \in Z, (j, osc) \mid I_{j,osc} \in P\}$ (a wavelet packet) an orthonormal basis of $L^2(R)$.

Also the CP structures offer optimal bases in terms of compression power and sparsity (Donoho *et al.*, 1998) and for dealing with non-stationary processes showing time-varying covariance operators (Mallat *et al.*, 1998). The building blocks are localized cosine functions, i.e., functions localized in time and forming smooth basis functions capable of representing dense covariance matrices in sparse form.

The CP transform has an advantage over the classic discrete cosine transform (DCT); while the latter defines an orthogonal transformation and thus maps a signal from the time to the frequency domain, it is not localized in time and thus unapt for handling non-stationary signals. Depending instead on taper functions that one may select, the cosine packets decay to zero within their compact support and because of this smoothness they overcome the limitations of DCT.

The classical DCT-II transform is defined as $g_k = \sqrt{\frac{2}{n}} s_k \sum_{i=0}^{n-1} f_{i+1} cos(\frac{(2i+1)k\pi}{2n})$, for $k = 0, 1, \ldots, n-1$, and with a scale factor s_k resulting 1 if $k \neq 0$ or n, and $\frac{1}{\sqrt{2}}$ if $k = 0$ or n. An inverse DCT is given by $f_{i+1} = \sqrt{\frac{2}{n}} \sum_{k=0}^{n-1} g_k s_k cos(\frac{(2i+1)k\pi}{2n})$. for $i = 0, 1, \ldots, n-1$.

5.3 Greedy Approximation

Given a random function $f \in H$ and a set D of vectors that densely span an Hilbert space H, one key problem is finding the best approximation of the unknown object by linearly combining a fixed number m of dictionary elements. The set D can be an orthonormal basis of H or can be redundant; both these cases bring advantages and disadvantages, but greedy algorithms can be used to provide an n-term approximation of f. A crucial issue is to what degree this approximation is sub-optimal (DeVore, 1998; Temlyakov, 2000).

The Matching Pursuit (MP) algorithm (Mallat and 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 libraries, whose signal representation is given by:

$$WP = \sum_{j,osc,k} w_{j,osc,k} W_{j,osc,k} + res_n \tag{35}$$

$$CP = \sum_{j,k} c_{j,k} C_{j,k} + res_n \tag{36}$$

In summary, the MP algorithm approximates a function with a sum of n elements, called atoms or atomic waveforms, which are indexed by 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 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 t 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:

$$f = \sum_{i=1}^{n} h_i H_{\gamma_i} + res_n \tag{37}$$

given the algorithmic steps:

1. INIZIALIZE WITH $res_0 = f$, AT I=1;
2. COMPUTE AT EACH ATOM H_γ THE PROJECTION $\mu_{\gamma,i} = \int res_{i-1}(t) H_\gamma(t) dt$;
3. FIND IN THE DICTIONARY THE INDEX WITH THE MAXIMUM PROJECTION,

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

   FROM THE ENERGY CONSERVATION EQUATION $argmax_{\gamma \in \Gamma} \mid \mu_{\gamma,i} \mid$;
4. WITH THE n^{th} MP COEFFICIENT h_n (OR $\mu_{\gamma_n,n}$) AND ATOM H_{γ_n}, COMPUTE THE UPDATED n^{th} RESIDUAL VIA:

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

5. REPEAT THE PROCEDURE FROM STEP 2, UNTIL $i \leq n$.

With $\mathcal{H}$ representing an Hilbert space, the function $f \in \mathcal{H}$ can thus be decomposed as $f =< f, g_{\gamma_0} > g_{\gamma_0} + Rf$, with f approximated in the g_{γ_0} direction, orthogonal to Rf, such that $\|f\|^2 =| < f, g_{\gamma_0} > |^2 + \|Rf\|^2$.

The latter is an energy conservation law which suggests that in order to minimize the $\|Rf\|$ term, it is required a choice of g_{γ_0} in the dictionary such that the inner product term is maximized (up to a certain optimality factor)[7].

6 Application

6.1 Data and Exploratory Analysis

The Nikkei stock return index series for the year 1990 and with observations collected at a very high frequency, i.e. every minute (1min), is available for empirical investigation. The total sample has 35,463 observations, with intra-daily trading prices covering the working week, and excluding holidays and weekends. A temporally aggregated time series of correspondent five-minute (5min) data is then formed from the original one by simply averaging the 1m-sampled time points. The aggregated sample consists of 7092 observations[8].

We compare the estimated spectrum with raw and smoothed periodograms to the denoised estimates obtained with a wavelet decomposition; in other words, the reconstructed spectrum is

[7]The choice of these atoms from the D dictionary occurs by choosing an index γ_0 based on a certain choice function conditioned on a set of indexes $\Gamma_0 \in \Gamma$.

[8]The experiments have been conducted with *S-Plus* and *Matlab*.

given by the application of a thresholding algorithm to the log periodogram $\Pi(w_k) = log \mid X(w_k) \mid$, where $X(w_k)$ is the discrete Fourier transform and $w_k = 2\pi k$ are the fundamental frequencies.

One can see the effect of transforming the data (Figure 1 top plots), smoothing the periodogram with triangular spectral windows of various widths (middle plots) and removing the noise (bottom plots).

It appears that the wavelet spectral estimator adapts well and further smooths the periodogram. No substantial boundary effects appear, given the large fluctuations at the highest frequencies; this means that the wavelet thresholding[9] works like a variable bandwith smoother and results quite robust.

Figures 2 and 3 report, respectively, the time-frequency tilings from the WP and the CP decompositions (A) and the top-100 largest coefficient approximation with the MP algorithm (B). The two dictionaries are partitioning time and frequency sequentially but in a different way; while WP divide the frequency domain in intervals which are then temporally partitioned, a reversed time-frequency space partitioning comes with CP.

In Figure 2 it is also reported a sub-sample performance after iterating 50, 100, 200 and 500 times (C-D, from top to bottom and from left to right); only for the WP case, as an example of how MP works through the iterations. These two groups of four plots refer to a sample splitting procedure into segments; this is one of the strategies which can be adopted to tackle non-stationarity in long time series.

The segment size choice here simply reflects the dyadic sample design, with a splitting rule which restricts the partition to sample sizes divisible by 2^J. At its simplest level of partitioning, it amounts to just considering two sub-samples, with observation ranges 1-3328 and 3329-7040, for the 5min series.

The plots suggest that MP works efficiently, due to its greedy nature, and 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.

6.2 More Experiments

One of 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 its capability of dealing efficiently with the so-called (Davis *et al.*, 1997) coherent structures compared to the dictionary noise components.

In volatility applications an issue is to control the behaviour of the transformed residue term after n approximation steps, i.e., the residual absolute and squared values are to be monitored since their autocorrelation functions give information about the conditional variance, and thus are of direct interest for the volatility modeling aspects.

The suggested *approximation + decomposition* method is below sketched, according to some steps:

- compute Matching Pursuit time-frequency tiling (with/without sample-splitting) and time-scale decomposition via WP and CP dictionaries;
- compute ICA in feature space selecting among the most informative (in terms of energy) resolution levels from the WP and CP signals;
- reduce the problem dimension accordingly (i.e., restrict the range of scales to be used in the atomic dictionaries) and re-start MP;

[9]The level-wise threshold is given by $\lambda_j = max(\pi\sqrt{\frac{log(n)}{3}}, log(2n)2^{\frac{-(j-1)}{4}})$ (Bruce and Gao, 1994).

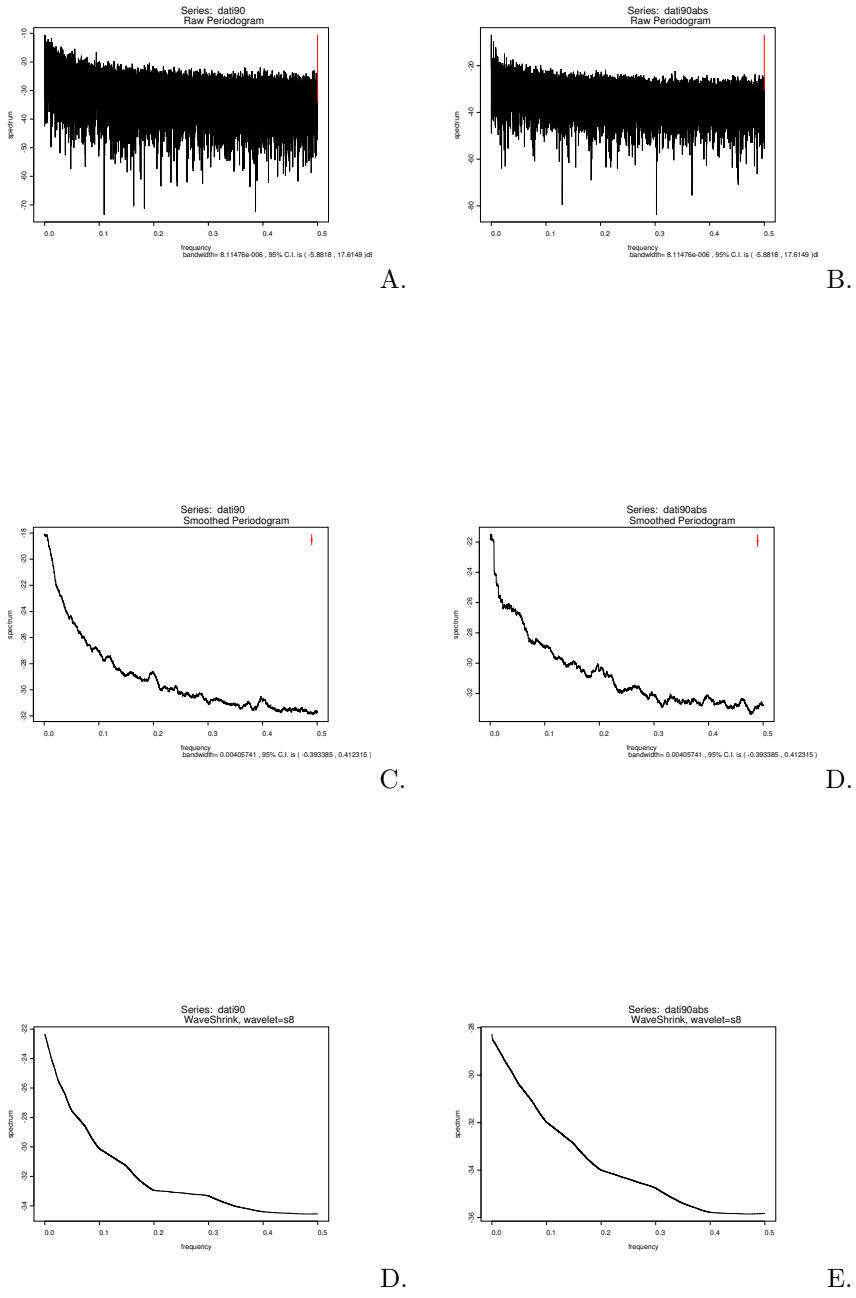


Figure 1: Top: raw periodogram. Middle: smoothed periodogram. Bottom: wavelet de-noised spectral estimate. Original 1m returns (left) and absolute values (right).

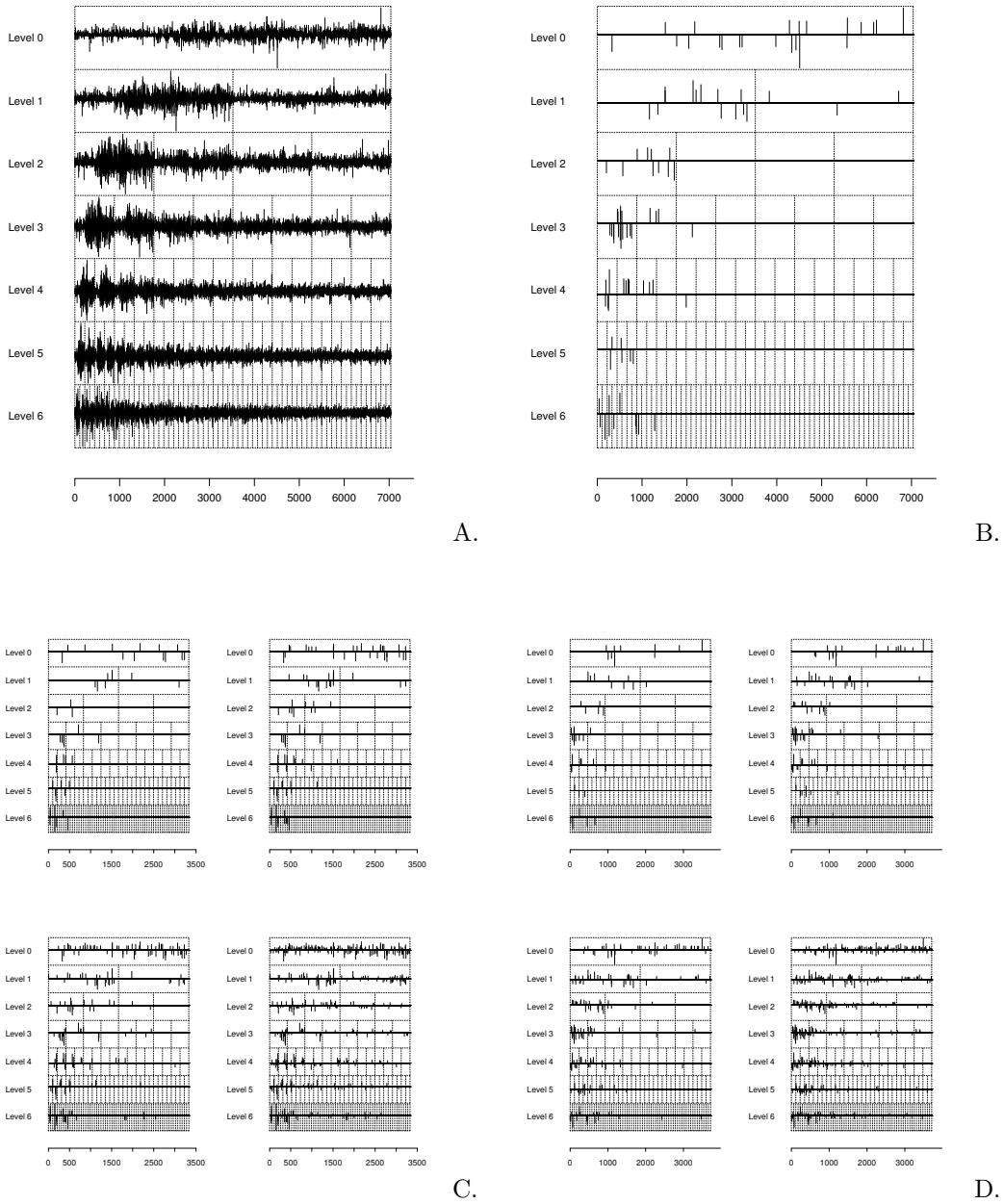
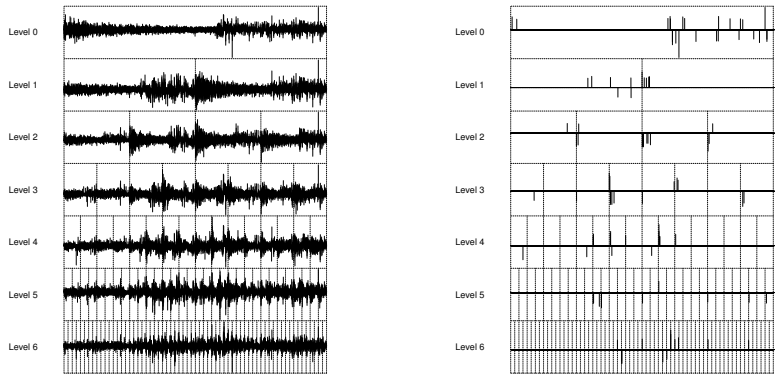


Figure 2: WP time-frequency tiling (A) and MP approximation with 100 atomic structures (B). MP progressive approximation power for 1st (C) and 2nd (D) sub-samples, with 50 (top left), 100 (top right), 200 (bottom left) and 500 atoms.



A. B.

Figure 3: CP time-frequency tiling (A) and MP approximation with 100 atomic structures (B).

- CONTROL THE FEATURE DETECTION POWER VIA THE AUTOCORRELATION FUNCTION AND THE MULTISCALING REGIMES VIA BOTH VARIANCE AND ENERGY PLOTS, FOR THE ATOMIC SEQUENCES AND THE EXTRACTED SOURCE SIGNALS.

The profile (Figure 4) and surface (Figure 5) plots show the across-scale variability of the WP and the CP sequences (A-C and B-D, respectively), together with that of the corresponding sources (E-G and F-H, respectively). With the sources there is higher variability and more heterogeneous behavior compared to the sequences.

Figure 6 shows auto- and cross-correlation functions computed for the highest two WP sequences (A) and correspondent sources (B); one may note that the sources improve both the autocorrelation structure and, in part, the cross-correlation structure.

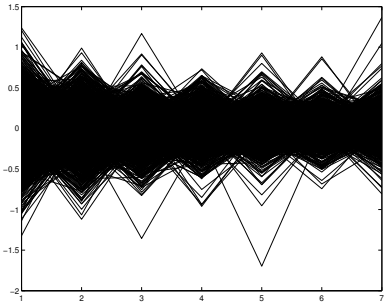
6.3 Dealing with Dependence and Multiscaling

By expanding a function over some basis, the dependence structure is learned through both the basis and the expansion coefficients; it has been shown (Abry *et al.*, 1998) that for long memory processes it is mostly up to the wavelet basis to capture dependence, leaving an almost uncorrelated sequence of expansion coefficients.

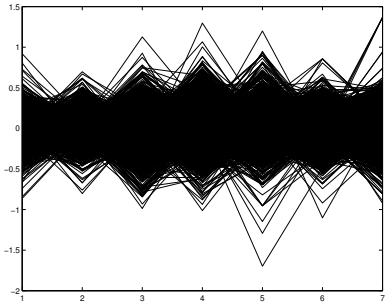
In a non-i.i.d. and non-stationary context, these aspects are even more emphasized and heteroscedasticity plays a major role, requiring a different treatment via resolution-wise thresholding. Now the information structure has to be accounted for by the correlation kernel, requiring a larger threshold compared to the orthogonal setting.

With regard to the projected reproducing kernel, the number of resolution levels for the chosen wavelet system characterizes the degree of correlation found through the transform in the wavelet coefficient domain. One of the most important properties of wavelets are the de-correlation and stationarizing power, appearing from the expansion coefficient sequences.

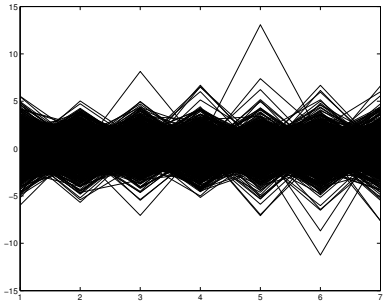
For a stochastic process X with strong dependence structure (Abry *et al.*, 2000), and given the wavelet coefficients $E[d_{j,k}] = 0$, the variance is:



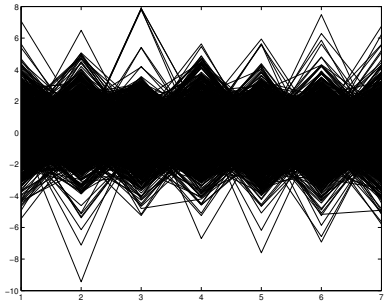
A.



B.

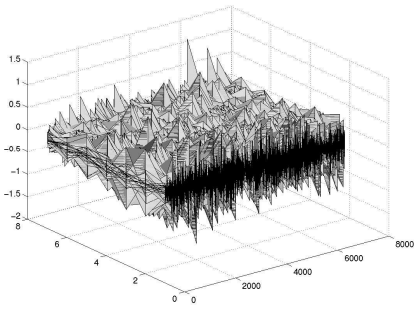


C.

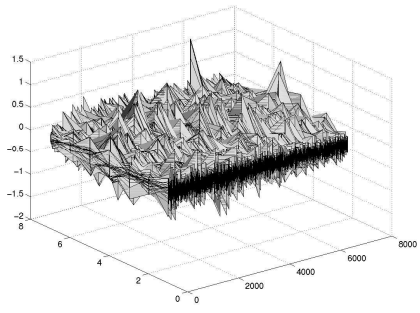


D.

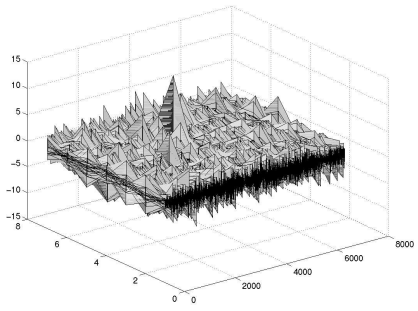
Figure 4: Resolution-wise variability for WP (left) and CP (right) sequences (profiles A-B) and sources (profiles C-D).



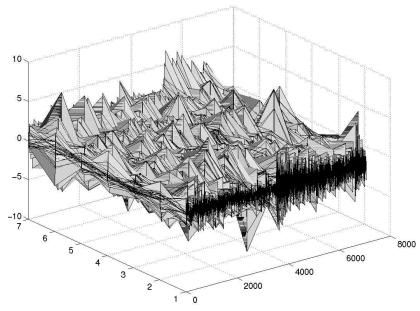
A.



B.

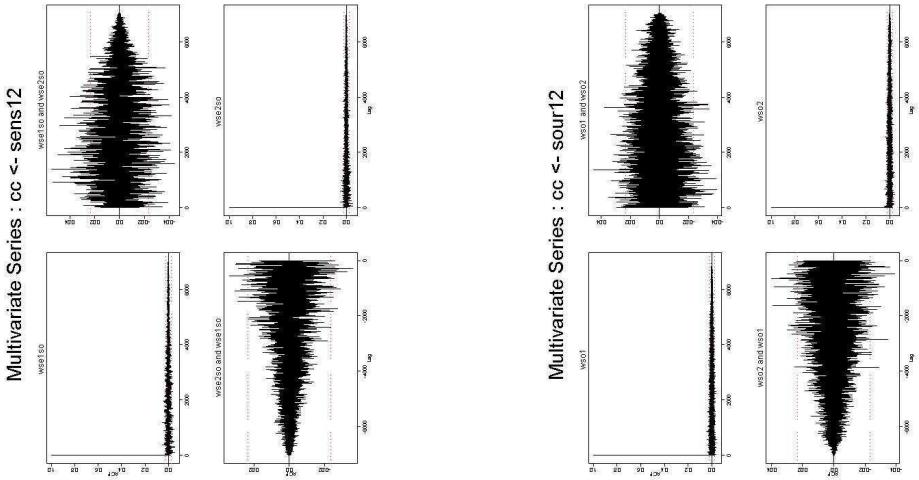


C.



D.

Figure 5: Resolution-wise variability for WP (left) and CP (right) sequences (surfaces A-B) and sources (surfaces C-D).



A. B.

Figure 6: Auto- and Cross-correlation functions for first and second resolution WP sequences (A). Same functions for corresponding WP sources (B).

$$E[d_{j,k}^2] = \int \Gamma_x(v) 2^j \mid \Psi_0(2^j v) \mid^2 dv \tag{38}$$

and represents a measure of the power spectrum $\Gamma_x(.)$ at frequencies $v_j = 2^{-j} v_0$, given the central frequency v_0 which depends on ψ_0, and with Ψ_0 the Fourier Transform of ψ_0.

With $\Gamma_x(v) \sim c_f \mid v \mid^{-\alpha}$ and $v \to 0$, i.e., under a power law behavior, then for $j \to +\infty$ and for $\alpha \in (0,1)$:

$$E[d_{j,k}^2] \sim 2^{j\alpha} c_f C(\alpha, \psi_0) \tag{39}$$

with $C(\alpha, \psi_0) = \int \mid v \mid^{-\alpha} \mid \Psi_0(v) \mid^2 dv$.

The covariance function of the wavelet coefficients is controlled by the number of vanishing moments M; when there is a sufficiently high number of them, they lead to high compression power, due to the fact that the finest coefficients are negligible for the smooth part of the function. In particular, the decay is much faster in the wavelet expansion coefficients domain than in that originated by long memory processes.

Note that a process is self-similar with Hurst parameter $H > 0$ if $X(st)$ and $s^H X(t)$ have identical distributions. With long range dependence, or asymptotic self-similarity, $\alpha = 2H - 1$ is the derived exponent.

The sequence $\{d_{j,k}\}_{k \in Z}$ of detail signals or wavelet expansion coefficients is a stationary process if the number of vanishing moments M satisfies a constraint, and the variance of $d_{j,k}$ shows scaling behaviour in a range of cut-off values $j_1 \leq j \leq j_2$ to be determined. The sequence no longer shows

long range dependence (LRD) but only short range dependence (SRD) when $M \geq \frac{\alpha}{2}$, and the higher M the shorter the correlation left, due to:

$$E[d_{j,k}d_{j,k'}] \approx \mid k - k' \mid^{\alpha-1-2M}, for \mid k - k' \mid \rightarrow +\infty \tag{40}$$

$$E[d_{j,k}d_{j',k'}] \approx \mid 2^{-j}k - 2^{-j'}k' \mid^{\alpha-1-2M}, for \mid 2^{-j}k - 2^{-j'}k' \mid \rightarrow +\infty \tag{41}$$

These assumptions don't rely on a Gaussian signal, and could be further idealized by assuming $E[d_{j,k}d_{j,k'}] = 0$ for $(j,k) \neq (j,k')$ and $E[d_{j,k}d_{j',k'}] = 0$ for $(j,k) \neq (j',k')$. One can then look at the variance as follows:

$$var(d_{j,k}) \approx 2^{j\alpha} \tag{42}$$

for $j \rightarrow \infty$. Thus, with an LRD process, the effect of the wavelet transform is clear, bringing back decorrelation or small SRD due to the control of non-stationarity and dependence through the parameter M.

Figure 7 shows the variance of the WP sequences and sources[10] computed at each resolution level, together with the log_2-energy plots (see Gilbert (2001) among others) measuring the average energy of the signal at each scale j, and along time and frequency coordinates, according to:

$$E_j = \frac{1}{N_j} \sum_{k=0}^{N_j-1} \mid WP_{j,k} \mid^2 \tag{43}$$

This quantity is to be considered a wavelet periodogram, and represents a useful spectral estimate computed along the number N_j of WP squared empirical coefficients available at each scale. Due to the redundancy of the WP transform, the presence of bias is inevitable, and how to handle this aspect is worth future investigation.

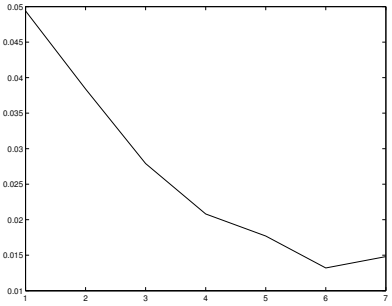
Despite this limitation, the results obtained are informative when comparing the log-energy plot obtained from the WP coefficient sequences with the corresponding variance plot. Compared to the behavior in the latter plot, where the shift at scales 4-5 is negligible, the linearly decreasing behavior coming from the highest scales is interrupted in the log-energy plot by an upward step when resolution level 4 appears, is followed by a further decrease at scale 5 and then a new jump occurs at level 6.

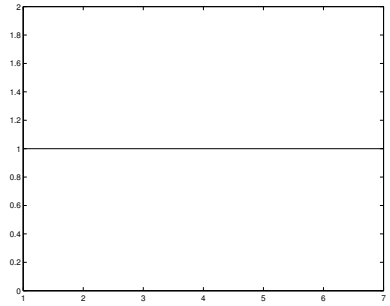
This behavior is in part expected, being the scaling effects usually observed only in a limited central range of scales, but it might be affected by bias too. However, it may also indicate the presence of multiscaling regimes, thus letting the nature of the underlying volatility process be responsible for the non-linear signatures characterizing the energy plot.

The source energy-plots offer an interesting support as a diagnostic tool. In particular, they might allow for an even sharper identification of regime thresholds at various scales. This because they are by construction independent, at least to a greater degree compared to the WP coefficient sequences, and thus one explains the behavior of plot D.

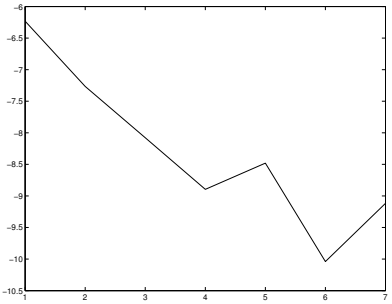
But one may observe that the two WP jumps at scale 4 and 6 are now emphasized (even if with different sign of the slope) by the sources, indicating a strong change of dynamics in that range of scales. This confirms what observed in Figure 4-C, where from scale 4 the variability of the sources intensifies, compared to that of the WP sequences which appear in Figure 4-A much more homogeneous.

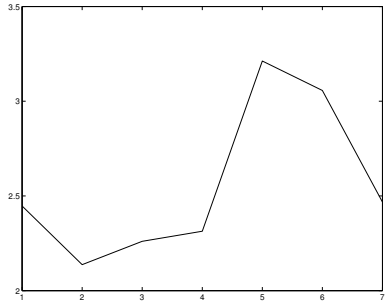
[10]They are normalized, hence the value 1. We limit the analysis to the WP case, due to the tiling structure of the two dictionaries.





A. B.





C. D.

Figure 7: WP scale-wise variance for WP (A) and sources (B); log_2-energy plots for WP (C) and sources (D).

The energy function refers to the second moment of the observed process at each resolution level, and thus represents a non-parametric unbiased estimator of the variance of the WP coefficients sequence.

The power law structure of the function $E(WP_{j,k}^2)$ usually suggests that E_j is that of a self-similar or of a long range dependent process, due to a log-linear relation with slope $2H-1$ or α, respectively.

The self similarity typical of financial series is here subject to the lens of the wavelet packet sequences; due to a power law behavior that results not homogeneous across all scales, it reveals itself only within a range of them.

7 Conclusions

One goal of this work was exploring multi-scaling cascade approximation systems that may be effective for solving inverse problems; the volatility recovery problem is one of them, and high-frequency stock index returns have been investigated so as to represent a volatility process through risk and return dynamics cast in a latent variable model.

The link which is established between risk and returns appears through a flexible functional relation that approximates a non-linear one; mixture and cascade dynamics can be thought to characterize this model, while the presence of regimes in the volatility structure is revealed by a scale-dependent behavior of ad-hoc diagnostic quantities such as log-energies.

The algorithmic learning strategy here adopted combines three main building blocks in an unified methodological framework aimed to perform sparse greedy approximation. While wavelet-like representations allow for multi-scale atomic decompositions of return realizations, a greedy approximation algorithm like the matching pursuit runs over atomic overcomplete dictionaries consisting of wavelet and cosine packets.

Statistical dimensionality reduction is then envisaged and built-in, as an independent component analysis step lets the greedy learner obtain near-optimal approximation with relatively few iterations. The special feature of the proposed approximation-decomposition algorithm is that it works sequentially in function and feature space, and allows for effective calibration and improvement of both volatility approximation and feature detection power.

Acknowledgment

The author wishes to thank anonymous referees for their careful reading of the manuscript and the fruitful comments and suggestions aimed to improve the paper.

References

[1] P. Abry, D. Veitch and P. Flandrin, Long range dependence: revisiting aggregation with wavelets. *J. Time Ser. An.*, **19(3)** 253-266 (1998).

[2] P. Abry, P. Flandrin, M.S. Taqqu and D. Veitch, Wavelets for the analysis, estimation and synthesis of scaling data. In *Self-similar network traffic and performance evaluation* (Eds. C. Park, W. Willinger, Wiley, New York.) 39-88 (2000).

[3] T. Andersen and T. Bollerslev, Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *Int. Ec. Rev.*, **39** 885-905 (1998).

[4] A. Antoniadis and C. Lavergne, Variance Function Estimation in Regression by Wavelet Methods. In *Wavelets and Statistics*, (Eds. A. Antoniadis and G. Oppenheim, Springer-Verlag, New York) 31-42 (1995).

[5] A. Arneodo, E. Bacry and J.F. Muzy, Random cascades on wavelet dyadic trees. *J. Math. Phys.*, **39(8)** 4142-4164 (1998).

[6] N. Aronszajn, Theory of Reproducing Kernels. *Trans. Am. Math. Soc.*, **686** 337-404 (1950).

[7] O.E. Barndorff-Nielsen and N. Shephard, Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion). *J. R. Statist. Soc. B*, **63** 167-241 (2001)

[8] T. Bollerslev, A generalized autoregressive conditional heteroskedasticity. *J. Econometr.*, **31** 307-327 (1986).

[9] E. Capobianco, Statistical Analysis of Financial Volatility by Wavelet Shrinkage. *Method. Comp. in Appl. Prob.*, **I(4)** 423-443(1999).

[10] E. Capobianco, Multiresolution Approximations for Volatility Processes. *Quant. Fin.*, **2** 91-110 (2002).

[11] E. Capobianco, Hammerstein System Representation of Financial Volatility Processes. *Eur. Phys. J. B*, **27(2)** 201-212 (2002).

[12] E. Capobianco, Independent Multiresolution Component Analysis and Matching Pursuit. *Comp. Statist. Data Anal.*, **42(3)** 385-402 (2003).

[13] E. Capobianco, Functional Approximation in Multi-scale Complex Systems. *Advances in Complex Systems*, **6(2)** 177-204 (2003).

[14] J. Cardoso, Source separation using higher order moments. *Proc. Int. Conf. Ac. Sp. Sig. Proc.*, 2109-2112 (1989).

[15] J. Cardoso and A. Souloumiac, Blind beamforming for non-Gaussian signals. *IEE Proc. F.*, **140(6)** 771-774 (1993).

[16] R. Carmona, W.L. Hwang and B. Torresani. *Practical Time-Frequency Analysis*. Academic Press, San Diego, 1998.

[17] S. Chen, D. Donoho and M.A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Rev.*, **43(1)** 129-159 (2001).

[18] P.K. Clark, A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices? *Econometr.*, **41** 3-32 (1973).

[19] R.R. Coifman, Y. Meyer and M.V. Wickerhauser, Wavelets analysis and signal processing. In *Wavelets and their Applications*, (Eds. B. Ruskai *et al.*, Jones and Barlett, Boston) 153-178 (1992).

[20] P. Comon, Independent Component Analysis - a new concept? *Sig. Proc.*, **36(3)** 287-314 (1994).

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

[22] I. Daubechies. *Ten Lectures on wavelets*. SIAM, Philadelphia, 1992.

[23] G. Davis, S. Mallat and M. Avellaneda, Greedy adaptive approximations. *Constr. Approx.*, **13(1)** 57-98 (1997).

[24] R. A. DeVore, Nonlinear Approximation. *Acta Num.*, 51-150 (1998).

[25] D. Donoho, Sparse Components of Images and Optimal Atomic Decompositions. *Constr. Approx.*, **17** 353-382 (2001).

[26] D. Donoho and I.M. Johnstone, Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika*, **81** 425-455 (1994).

[27] D. Donoho and I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage. *JASA*, **90** 1200-1224 (1995).

[28] D. Donoho and I.M. Johnstone, Minimax Estimation via Wavelet Shrinkage. *Ann. Statist.*, **26** 879-921 (1998).

[29] D. Donoho, S. Mallat and R. von Sachs, Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods. Tech. Rep. 1998-517, Dep. Statistics, Stanford University, Stanford, 1998.

[30] J. Durbin and S.J. Koopman, Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc.* B, **62** 3-56 (2000).

[31] R.F. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometr.*, **50** 987-1008 (1992).

[32] E. Ghysels, A.C. Harvey and E. Renault, Stochastic Volatility. In Handbook of Statistics 14, (Eds. G.S. Maddala and C.R. Rao, Elsevier Science, Amsterdam), 118-191 (1996).

[33] A.C. Gilbert, Multiscale Analysis and Data Networks. *Appl. Comp. Harm. An.*, **10** 185-202 (2001).

[34] W. Greblicki, Continuous-Time Hammerstein System Identification. *IEEE Tr. Aut. Contr.*, **45(6)** 1232-1236 (2000).

[35] U. Grenander. *Abstract Inference*. Wiley, New York, 1981.

[36] Z. Hasiewicz, Non-parametric estimation of non-linearity in a cascade time-series by multiscale approximation. *Sig. Proc.*, **81** 791-807 (2001).

[37] I.M. Johnstone, Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sin.*, **9** 51-83 (1999).

[38] I.M. Johnstone and B.W. Silverman, Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B, **59** 319-351 (1997).

[39] I. Karatzas and E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, 1998.

[40] M.S. Lewicki and T.J. Sejnowski, Learning Overcomplete Representations. *Neur. Comp.*, **12(2)** 337-365 (2000).

[41] S. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L_2(R)$. *Trans. Am. Math. Soc.*, **315** 69-87 (1989).

[42] S. Mallat, G. Papanicolaou and Z. Zhang, Adaptive Covariance Estimation of Locally Stationary Processes, *The Ann. of Statist.*, **26(1)** 1-47 (1998).

[43] S. Mallat and Z. Zhang, Matching Pursuit with time frequency dictionaries. *IEEE Tr. Sig. Proc.*, **41** 3397-3415 (1993).

[44] S. Mallat and S. Zhong, Characterization of signals from multiscale edges. *IEEE Trans. Patt. An. Mach. Int.*, **14(7)** 710-732 (1992).

[45] I. Meyer. *Wavelets: algorithms and applications*. SIAM, Philadelphia, 1993.

[46] G.P. Nason and R. von Sachs, Wavelets in Time Series Analysis. *Phil. Trans. R. Soc. Lond. A*, **357** 2511-2526 (1999).

[47] G.P. Nason, R. von Sachs and G. Kroisandt, Wavelet Processes and Adaptive Estimation of the Evolutionary Wavelet Spectrum. *J. R. Statist. Soc. B*, **62** 271-292 (2000).

[48] E. Parzen, An approach to time series analysis. *Ann. Math. Stat.*, **32** 951-989 (1961).

[49] E. Parzen, Statistical Inference on Time Series by RKHS methods. In *12th Biennial Sem. Can. Math. Congr. Proc.* (ed R. Pyke) and Tech. Rep. n. 14, Stanford University (US), 1970.

[50] J.C. Ralston, A.M. Zoubir and B. Boashash, Identification of a Class of Nonlinear Systems under Stationary Non-Gaussian Excitation. *IEEE Tr. Sig. Proc.*, **45(3)** 719-735 (1997).

[51] V.N. Temlyakov, Weak Greedy Algorithms. *Adv. Comp. Math.*, **12(2,3)** 213-227 (2000).

[52] M. Zibulewsky and B.A. Pearlmutter, Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neur. Comp.*, **13(4)** 863-882 (2001).