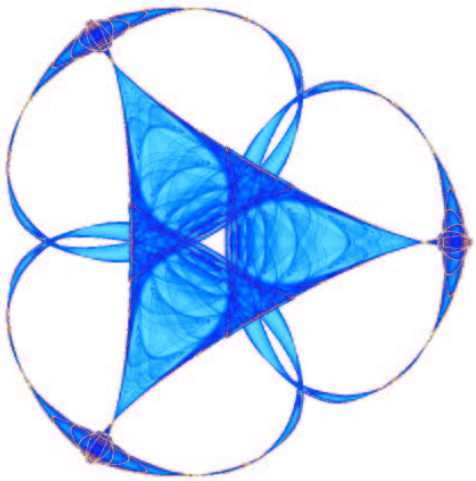
FRAMEABLE NON-STATIONARY PROCESSES AND VOLATILITY APPLICATIONS

By

Enrico Capobianco

IMA Preprint Series # 2011

(December 2004)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Frameable Non-stationary Processes and Volatility Applications

Enrico Capobianco
Biomedical Engineering Department,
Boston University, 44 Cummington St., Boston, MA 02215 USA
ecapob@bu.edu

December 4, 2004

Abstract

A crucial goal in many experimental fields and applications is achieving sparse signal approximations for the unknown signals or functions under investigation. This fact allows to deal with few significant structures for reconstructing signals from noisy measurements or recovering functions from indirect observations. We describe and implement approximation and smoothing procedures for volatility processes that can be represented by frames, particularly wavelet frames, and pursue these goals by using dictionaries of functions with adaptive degree of approximation power. Volatility is unobservable and underlying the realizations of stochastic processes that are non-i.i.d., covariance non-stationary, self-similar and non-Gaussian; thus, its features result successfully detected and its dynamics well approximated only in limited time ranges and for clusters of bounded variability. Both jumps and switching regimes are usually observed though, suggesting that either oversmoothing or de-volatilization may easily occur when using standard and non-adaptive volatility models. Our methodological proposal combines wavelet-based frame decompositions with blind source separation techniques, and uses greedy de-noisers and feature learners.

Keywords: *Frames and Wavelets; Blind Source Separation; Sparse Approximation; Greedy Algorithm; Volatility Smoothing.*

MSC2000: 60G35, 62M10, 65T60, 93E03.

1 Introduction

Volatility processes have been for almost two decades a cross-disciplinary research subject and have suggested many challenging problems to financial mathematicians and statisticians. The main interest of this work too is in volatility, particularly the approximation of its dynamics and its statistical modelling. It is useful to start from the following general classification.

Volatility can be specified as conditionally dependent on past squared returns and own lags, thus representing the class of *Generalized Autoregressive Conditionally Heteroscedastic* (GARCH) processes [29, 7], or as a stochastically indipendent process, characterized by markovian structure and a noise source independent from the disturbance term in the conditional mean equation, thus referring to *Stochastic Volatility* (SV) processes [32].

The main stylized facts about volatility models appear from many empirical studies [46], and among them the most important are (1) heavy-tailed (leptokurtic) marginal return distributions; (2) volatility clustering, and thus tail dependence; (3) second-order dependence, visible in absolute and squared transformed returns; (4) long memory and covariance nonstationarity.

The GARCH and SV classes are often selected according to different goals, since they refer to different information sets. In the first model class there is observation-driven information, with $y_t \mid Y_{t-1} \sim N(0, \sigma_t^2)$, for $\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \ldots + \alpha_p y_{t-p}^2$ (i.e., ARCH(p) case); thus, the information set $\mathcal{F}_t$ for the index retun process is formed by past squared observations up to time *t-p*, i.e., $\sigma\{Y_s^2 : s \leq t-p\}$.

In the other class of models this is not possible, since they are driven by parameters and include both observable and unobservable variables; the index return process is distributed according to $y_t \mid h_t \sim N(0, exp(h_t))$, and the volatility is specified as $h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t$, $\eta_t \sim NID(0, \sigma_\eta^2)$. With η_t Gaussian, h_t is autoregressive of order one, and covariance stationarity follows if $\mid \gamma_1 \mid < 1$. Then $\mu_h = E(h_t) = \frac{\gamma_0}{1-\gamma_1}$, $\sigma_h^2 = var(h_t) = \frac{\sigma_\eta^2}{1-\gamma_1}$ and the kurtosis is $\frac{E(y_t^4)}{(\sigma_{y^2})^2} = 3exp(\sigma_h^2) \geq 3$, resulting in fatter tails than the Gaussian ones.

Non-parametric and semi-parametric approaches have been introduced with the aim of relaxing assumptions (see for instance the seminal work of [30]) on the error probability density of regression models, or on the unobservable volatility function itself. Usually one may end up with iteratively smoothing the unspecified function, as in log-transformed multiplicative models [33] where backfitting estimation procedures [34] are adopted.

In these cases, the flexibility allowed so to account for the unrestricted structure in the volatility function may be not sufficient when dealing with stochastic volatility, for then the consistency of estimation procedures may fail. This fact occurs when a certain orthogonality is built in the model, which relies on the statistical independence between the information sets to which the volatility and the conditional mean dynamics refer.

The procedures that we suggest and implement are recursive in nature, being this an important condition for building effective stochastic dynamics models, and are designed for operating under both orthogonal and non-orthogonal conditions. The reference framework is that of non-orthonormal and redundant frame-based systems, together with that of orthonormal systems built on atomic dictionaries o approximating functions.

It has been shown in many studies (see also [20] for an excellent review) that from such

systems, non-linear estimators can be built and result effective and computationally fast. For inhomogeneous function classes they result superior to linear estimators employed by backfitting methods with some kind of smoother. We illustrate how sparsified frame-based volatility models may be investigated by a novel combination of greedy approximation and space dimensionality reduction techniques that lead to near-optimal volatility smoothing and feature detection.

The paper is organized as follows: Section 2 reviews frames, while Section 3 and 4 illustrate, respectively, multiresolution and greedy approximation techniques. Section 5 suggests an estimation technique based on blind source separation analysis and reports experiments on modeling and smoothing volatility from a stock index return series. The conclusions are in Section 6.

2 Frames

A general way of approximating families of functions belonging to general spaces refers to *frames* [27, 19], which represent redundant sets of approximating vectors. Throughout the paper, volatility processes will be considered to be non-stationary or piecewise stationary, as in local stationary processes [18, 47]; thus, their realizations belong to spaces that allow for spatial inhomogeneity and a certain complexity of dependence structure.

Together with the variance positivity constraint, a local boundedness condition for the volatility paths is also assumed; this allows for preventing the volatility function from explosive behavior. We thus consider frameable volatility functions those ones that can be represented by frames or similarly derived systems.

Frame components are not linearly independent, but despite this aspect which seems to penalize them from a computational standpoint, there are advantages in using frames since they lead to numerically stable and robust-to-noise reconstruction algorithms while also allowing for increased feature detection power, due to their flexibility.

A formal specification of frames requires the presence of a system $\{\gamma_k\}_{k=1}^{M}$ and bounds A and B such that:

$$A \mid\mid x \mid\mid^2 \leq \sum_{k=1}^{M} |< x, \gamma_k >|^2 \leq B \mid\mid x \mid\mid^2 \tag{1}$$

$\forall x \in R^N$.

A frame operator associated to them is F such that:

$$[Fx]_{k=1}^{M} =< x, \gamma_k > \tag{2}$$

Noting that the redundancy of the frame comes from $M \geq N$, and it is measured by the ratio M/N, the frame operator, when multiplied by its transpose F^*, shows two properties:

- F^*F is invertible and A^{-1} and B^{-1} are its bounds;
- a dual frame $\{\tilde{\gamma_k}\}_{k=1}^{M}$ is defined such that:

$$\tilde{\gamma_k} = (F^*F)^{-1}\gamma_k \tag{3}$$

where now the associated dual operator is:

$$\tilde{F} = F(F^*F)^{-1} \tag{4}$$

A tight frame is defined when $A = B$, while an orthonormal basis requires a value of 1 for the bounds; this allows for an improved reconstrution power, as in this case F^*F is diagonal and $col(F)$ are orthogonal, thus suggesting that a basis can be formed. Another aspect of interest is given by the reconstruction step, where an inverse or pseudoinverse of the F operator is required.

The standard formula is delivered by:

$$F^- = (F^*F)^{-1}F^* \tag{5}$$

but it can be shown that $F^- = \tilde{F}^*$, which means that one choice for the pseudoinverse is provided by the transpose of the operator associated to the dual frames, and this leads to a reconstruction formula which depends on this last operator as follows:

$$x = \tilde{F}^*Fx = \sum_{k=1}^{M} < x, \gamma_k > \tilde{\gamma_k} \tag{6}$$

This linear reconstruction leads with noisy signals to a change in the previous expansion so to include the effects of noise according to:

$$\sum_{k=1}^{M} [< x, \gamma_k > \ + \epsilon_k] \ \tilde{\gamma_k} \tag{7}$$

given the noise ϵ.

Note that a generalized reproducing kernel [41] is found whenever we have a vector $x_n \subset x$ (i.e., a subset of the frame coefficients $[Fx]_{k=1}^{M} =< x, \gamma_k >$) for which we compute the denoising projection:

$$Px_n = \sum_k x_k < \tilde{\gamma_k}, \gamma_k > \tag{8}$$

Since for a sequence of frame coefficients it holds that $x = Px$, our previous relation is valid for x too, due to the kernel $< \tilde{\gamma_k}, \gamma_k >$. The mean square error (MSE) is found as:

$$\frac{1}{N} E \mid\mid x - \hat{x} \mid\mid^2 = \frac{1}{N}\sigma^2 \sum_{k=1}^{M} \mid\mid \tilde{\gamma_k} \mid\mid^2 \tag{9}$$

with σ^2 the variance of the noise term. Given the definition of the dual frames, and replacing in the MSE formula, one may achieve some optimality results, as when an orthogonal matrix can be used a sequence of uniform frames is obtained which asymptotically approaches a tight frame. And it is precisely for this class of frames that the MSE results optimal, i.e., minimum.

The MSE computed with frames is of course related to sparsity and coherence measures; we might expect that more sparsity and coherence bring MSE closer to its minimum, due to the fact that the contribution coming from informative and/or correlated structures dominates that related to the noise, and a better performance in terms of minimizing errors is achieved by greedy approximation algorithms. These aspects will be introduced afterwards, while now special cases are illustrated.

3 Multiresolution Feature Learning

3.1 Wavelet Frames

Wavelet frames are constructed by sampling from the *Continuous Wavelet Transform* (WT^c) [19] over time and scale coordinates. 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 WT^c is:

$$WT^c(f)_{jk} =< f, \psi_{jk} >=\mid j \mid^{-\frac{1}{2}} \int f(t)\psi(\frac{t-k}{j})dt \tag{10}$$

The function f can be recovered from the following reconstruction formula:

$$f = c_\psi^{-1} \int \int WT^c(f)_{jk}\psi_{jk}\frac{djdk}{j^2} \tag{11}$$

and this comes from the "resolution of identity formula":

$$\forall f, g \in L^2(R), \quad \int \int WT^c(f)_{jk}\widehat{WT}^c(g)_{jk}\frac{dkdj}{j^2} = c_\psi < f, g > \tag{12}$$

Given a constant $c_\psi < \infty$ and integration ranging from $-\infty$ to ∞, the WT^c maps f into an Hilbert space, and its image, say $L^2_{WT^c}(R)$, is a closed subspace and a *Reproducing Kernel Hilbert Space* too, since from the resolution formula it is enough to replace g with ψ so to get:

$$K(j, k, j', k') =< \widehat{WT}^c(g)_{jk} > = < \psi_{j'k'}, \psi_{jk} > \tag{13}$$

More generally, given a *scaling function* 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 $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$, one obtains differences among approximations computed at successively coarser resolution levels.

Thus, a so-called *Multiresolution Analysis* (MRA) [40, 19] is obtained, i.e. *a sequence of closed subspaces 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* $f \in V_j \iff f(2^j.) \in V_0$. 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, Sobolev, Holder, for instance, and in general all Besov and Triebel spaces [45].

Generally speaking, with a *Discrete Wavelet Transform* (DWT) a map $f \to w$ from the signal domain to the wavelet coefficient domain is obtained; 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:

$$f(x) = \sum_k c_{j_0 k}\phi_{j_0 k}(x) + \sum_{j>j_0}\sum_k d_{jk}\psi_{jk}(x) \tag{14}$$

where $\phi_{j_0 k}$ is associated with the corresponding coarse resolution coefficients $c_{j_0 k}$ and d_{jk} are the detail coefficients, i.e., $c_{jk} = \int f(x)\phi_{jk}(x)dx$ and $d_{jk} = \int f(x)\psi_{jk}(x)dx$. In short, the first term of the right hand side of (1) is the projection of f onto the coarse approximating space V_{j_0} while the second term represents the cumulated details.

3.2 De-noising and De-correlation

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:

$$\tilde{d}_{jk} = d_{jk} + \epsilon_t \tag{15}$$

The *wavelet shrinkage principle* [23, 24, 25] applies a thresholding strategy which yields de-noising of the observed data; it operates by shrinking wavelets coefficients to 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 the hidden dependence, de-noising can be useful for spatially heterogeneous signals.

Financial time series are realizations of non-stationary and non-Gaussian stochastic processes; a reason why wavelet systems could be effectively used when dealing with these signals, concerns the ability of wavelets to de-correlate along time and across scales. The de-correlation effect of the wavelet coefficients is one of the main properties that wavelet transforms bring into the analysis [36, 2]; thus, from a structure with long range dependence (LRD), either short range dependence (SRD) or de-correlation are found, when looking at wavelet coefficients at each scale 2^j or resolution level j.

Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, consider a stochastic process and one of its realizations observed with strong dependence structure; following [1], and with a slight variation in the notation compared to before due to the frequency (scale) characterization, let $E[d_x(j,k)] = 0$ and the variance be:

$$E[d_x(j,k)^2] = \int \Gamma_x(v) 2^j \mid \Psi_0(2^j v) \mid^2 dv \tag{16}$$

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

Given $\Gamma_x(v) \sim c_f \mid v \mid^{-\alpha}$ and $v \to 0$, then:

$$E[d_x(j,k)^2] \sim 2^{j\alpha} c_f C(\alpha, \psi_0) \tag{17}$$

for $j \to \infty$, and

$$C(\alpha, \psi_0) = \int \mid v \mid^{-\alpha} \mid \Psi_0(v) \mid^2 dv \tag{18}$$

for $\alpha \in (0,1)$.

One can then look at the variance law as follows:

$$var(d_x(j,k)) \approx 2^{j\alpha} \tag{19}$$

for $j \to \infty$.

The decay of the covariance function is much faster in the wavelet expansion coefficients domain than in the domain originated by long memory processes. The covariance function of the wavelet coefficients is controlled by M, the number of vanishing moments; when they are present in sufficiently high number they lead to high compression power.

The sequence of detail signals or wavelet expansion coefficients is a stationary process if the number of vanishing moments M satisfies the constraint that the variance of $d_x(j,k)$ shows scaling behaviour in a range of cut-off values $j_1 \leq j \leq j_2$ which has to be determined. The sequence no longer shows LRD but only SRD when $M \geq \frac{\alpha}{2}$, and the higher M the shorter the correlation left, due to:

$$E[d_x(j,k) d_x(j,k') \approx \mid k - k' \mid^{\alpha - 1 - 2M}, for \mid k - k' \mid \to \infty \tag{20}$$

These assumptions don't rely on a Gaussian signal, and could be further idealized by assuming $E[d_x(j,k) d_x(j,k') = 0$ for $(j,k) \neq (j,k')$.

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 M parameter. Across scales, instead, a certain degree of independence is obtained, so that the detail series might individually contribute to different information content in terms of frequency.

With regard to non-stationarity, an almost natural condition of financial time series, especially when measured at very high frequencies; wavelets stationarize the data when they are observed in their transformed wavelet coefficients, and they enable this change in a resolution-wise fashion. As suggested by [15] $X(t)$ is a stationary process if and only if it is stationary at all levels of resolutions:

Definition 1: k-stationarity
Stationarity at the kth level of resolution if $\forall n \geq 1, t_1, \ldots, t_n \in T$ and $l \in Z$ holds when:

$$[X(t_1 + 2^{-k}l), \ldots, X(t_n + 2^{-k})] \stackrel{d}{=} [X(t_1, \ldots, X(t_n)] \tag{21}$$

3.3 Multiscale Decomposition

We face two problems, when approximating the volatility function and estimating the model parameters involved: the role of smoothness and the presence of noise. Following [5], we might rely on quadratic information from the data, leading to non-negative estimators of the following kind:

$$\tilde{\sigma}^2(t) = \sum_i \alpha_i r_i^2(t) \tag{22}$$

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 this time through inner products:

$$\sigma_W^2(t) = \sum_k < \sigma^2(t), \phi_{j0,k} > \phi_{j0,k}(t) + \sum_{j>j_0} \sum_k < \sigma^2(t), \psi_{j,k} > \psi_{j,k}(t) \tag{23}$$

where a smooth part is combined with a cumulated sequence of details obtained at different scales.

We can then apply the same decomposition to $\tilde{\sigma}^2$, being σ_2 unobservable, and obtain a perturbed version of the previous approximation:

$$\hat{\sigma}^2(t) = \tilde{\sigma}_W^2(t) + \epsilon(t) \tag{24}$$

where $\epsilon = W\xi$ represents a wavelet transformed disturbance.

In order to sparsify the estimator, we can apply noise shrinkage techniques or use function expansions in different bases or overcomplete dictionaries.

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 [31].

3.4 Overcomplete Representations

Function dictionaries are collections of parameterized waveforms or atoms [14] ; they are available for approximating many classes of functions, formed directly from a particular family 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 represent remaining dictionary structures, thus originating a non-unique signal decomposition.

Wavelet packets (WP) [16] are one example of overcomplete representations; they represent an extension of the wavelet transform and allows for better adaptation due to an oscillation index f which delivers a richer combination of functions.

Given the admissibility condition $\int_{-\infty}^{+\infty} W_0(t)dt = 1$, $\forall (j,k) \in Z^2$, following [38] we have:

$$2^{-\frac{1}{2}} W_{2f}(\frac{t}{2} - k) = \sum_{i=-\infty}^{\infty} h_{i-2k} W_f(t-i) \tag{25}$$

where f relates to the frequency and h to the low-pass impulse response of a quadrature mirror filter. Also the following holds:

$$2^{-\frac{1}{2}} W_{2f+1}(\frac{t}{2} - k) = \sum_{n=-\infty}^{\infty} g_{n-2k} W_f(t-n) \tag{26}$$

where g is this time 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:

$$\{2^{-\frac{j}{2}} W_f(2^{-j}t - k), k \in Z, (j,f) \mid I_{j,f} \in P\} \tag{27}$$

One thus obtains a better domain compared to simple wavelets for selecting a basis to represent the signal and can always select an orthogonal wavelet transform by changing the partition P and defining $w_0 = \phi(t)$ and $W_f = \psi$, from the so-called WP transform (WPT).

Correspondingly, *Cosine Packets* (CP) and the related transform (CPT) suggest optimal bases in terms of compression power and sparsity [26], and optimal bases for dealing with non-stationary processes with time-varying covariance operators [43].

The building blocks in CP are localized cosine functions, i.e., localized in time and forming smooth basis functions, and they form almost eigenvectors for certain classes of non-stationary processes, and thus almost diagonal operators for approximating the covariance function.

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.

The *DCT-II* transform is defined as:

$$g_k = \sqrt{\frac{2}{n}} s_k \sum_{i=0}^{n-1} f_{i+1} cos(\frac{(2i+1)k\pi}{2n}) \tag{28}$$

for $k = 0, 1, \ldots, n-1$, and 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 instead given by:

$$f_{i+1} = \sqrt{\frac{2}{n}} \sum_{k=0}^{n-1} g_k s_k cos(\frac{(2i+1)k\pi}{2n}) \tag{29}$$

for $i = 0, 1, \ldots, n-1$.

4 Greedy Algorithms

Optimal algorithms often require adaptive signal approximation techniques based on sparse representations. Sparsity refers to the possibility of considering only few elements of a dictionary of approximating functions selected among a redundant set. In this way, by restricting the search to a sub-set of the original atomic dictionary, one may combine fast convergence through computing less inner products.

The MP algorithm [42] 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

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

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 indexed by 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 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:

$$f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + res_n(t) \tag{30}$$

and following the procedure:

1. *inizialize with* $res_0(t) = f(t)$, *at i=1;*

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

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* n^{th} *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 until* $i \leq n$*.*

With $\mathcal{H}$ representing an Hilbert Space, the function $f \in H$ 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 $\|f\|^2 =| < f, g_{\gamma_0} > |^2 + \|Rf\|^2$.

Thus, to minimize 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 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 $\Gamma_0 \in \Gamma$.

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 it is capable of dealing efficiently with the so-called *coherent structures* compared to the *dictionary noise* components [21].

5 Non-parametric Estimation

5.1 Data

Figure 1 reports one-day 1m returns compared to the 5m aggregated values, together with their absolute and squared transforms. We deal with stock returns observed at very high frequencies from the Nikkei 225 composite index; the data were collected minute by minute and refer to a certain market activity year, 1990.

We have approximately 35,000 observed values, covering intra-day market activity and excluding holidays and weekends; we have transformed the prices in financial returns by taking, as usual, the logarithms of ratios of consecutive time point prices.

We have also generated a temporally aggregated (every five minutes) series, which basically smooths the original series, at the price of losing high frequency content. More than 7,000 5-minute observations remain available to conduct a compared analysis.

We note a self-similar behavior and the typical function shape conditioned to the different intensity of intra-day activity hours according to usual market technical phases. In defining *self-similarity*, addressed as the property of *self-affinity* by [44], from [6] we have the following definition:

Definition 2: self-similarity

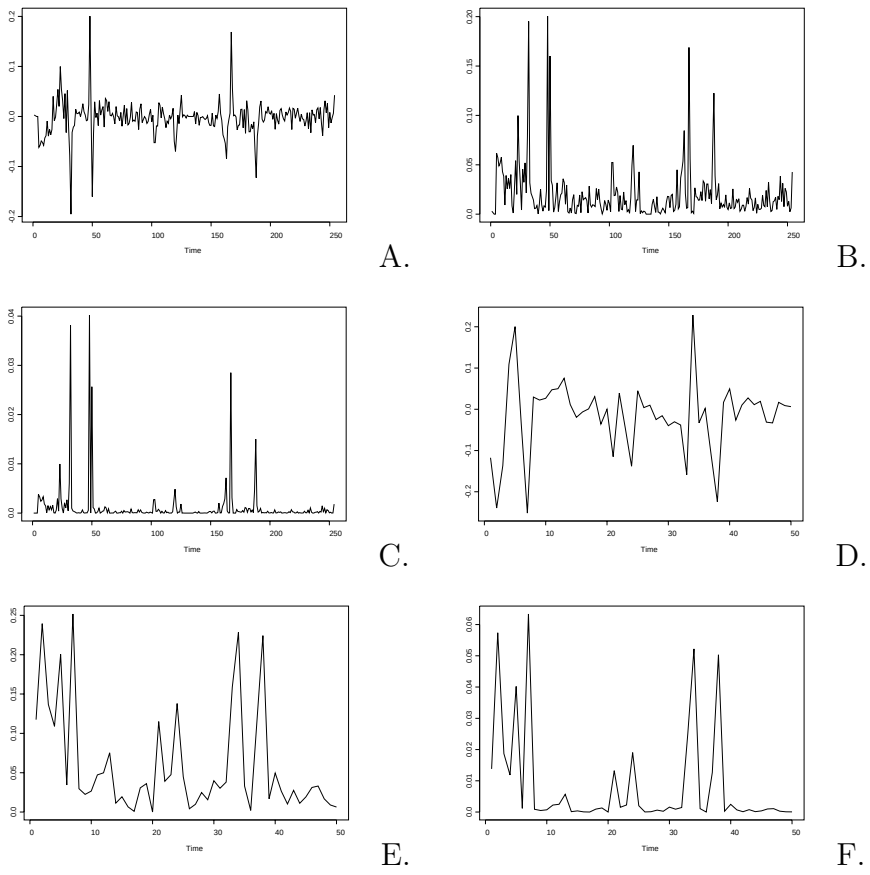


Figure 1: A) Raw 1m returns. B) Absolute 1m returns. C) Squared 1m returns. D) Raw 5m returns. E) Absolute 5m returns. F) Squared 5m returns.

Given $\alpha \in (0,1)$, $\gamma > 0$, $f : R^d \to R$ and $\bar{x} \in R^d$, a local re-normalization operator family, $R^{\gamma}_{\alpha,\bar{x}}$ can be constructed such that:

$$R^{\gamma}_{\alpha,\bar{x}} f(x) = \frac{1}{\gamma^{\alpha}}[f(\bar{x} + \gamma x) - f(\bar{x})] \quad \forall x \in R^d \tag{31}$$

Thus, for instance, a Gaussian process X defined on a probability space $(\Omega, \mathcal{F}, P)$, is a self-similar process of degree α if $[R^{\gamma}_{\alpha,\bar{x}} X \stackrel{\mathrm{d}}{=} X], \forall \gamma \in R^+$ and $\forall \bar{x} \in R^d$.

As a remark, working with just one long realization of the underlying return process means accepting the limitations that necessarily follow with regard to asymptotic inference, but at the same time represents a *de facto* typical situation in non-experimental contexts, where the suggested techniques might be used.

Dealing with non-stationarity is imposed by real circumstances; the observed series is subject to regime changes and external factors whose impact on the dynamics of returns and structure of volatility is undoubtedly relevant. The observed index return time series and the underlying volatility process represent a challenging context for approximation and estimation techniques.

5.2 Iterative Smoother

Instead of using very noisy squared return sequences, we may run the MP greedy algorithm directly over the returns and then look at the computed absolute and squared residuals for investigating the volatility features.

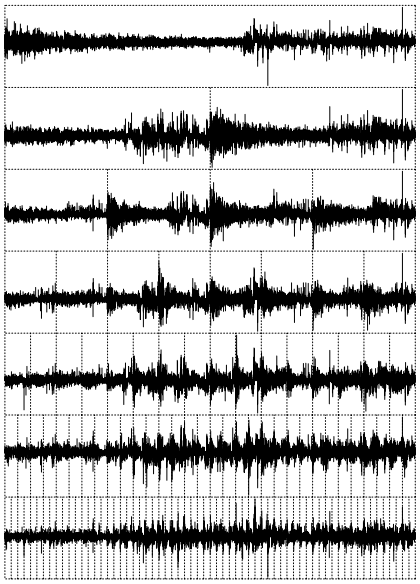
The corresponding autocorrelation function delivers a diagnostic measure of the feature detection power of the MP algorithm run over WP and CP dictionaries. Thus, in our experiments, a main issue is to control the behavior of the transformed MP residual terms after n approximation steps.

We have examined (see Figure 2) wavelet and cosine packets, whose time-frequency partitioning role is complementary. While cosine packets build a partition in the time domain and then run over each interval further frequency segmentation, wavelet packets work the other way around, by first segmenting along frequencies and then along time time. Thus, the former face frequency inhomogeneous smoothness while the latter deal with time inhomogeneity [28].

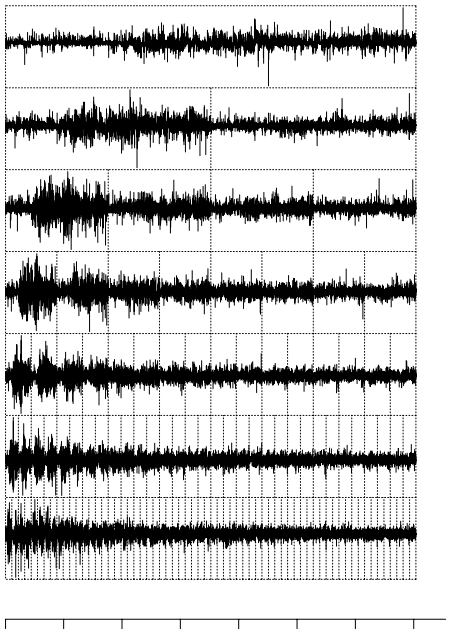
The within-block coefficients of the WP and CP structures describe the contribution at each resolution level of both time and frequency components in representing the signal features under a varying oscillation index varying from 0 to $2^J - 1$ (right-wise).

The WP Table presents sets of coefficients stored in sequency order according to an increasing oscillation index; at level 0 the original signal is represented and at level 1 the two subsets $w_{1,0}$ and $w_{1,1}$ have scale 2, corresponding to c_1 and d_1 obtained with the DWT. The CP table presents instead blocks ordered by time and the coefficients within the blocks are ordered by frequency.

The way these plots should be interpreted suggests that 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. For the CP table, the high frequency part of the signal is now expected on the left side, while the low frequency behavior appears from the right side.



A.



B.

Figure 2: CP table (A) and WP table (B) with signal segmentation level-by-level.

We investigate the performance of the MP algorithm when is applied on an *ad hoc* restricted and selected dictionary, based on a certain range of most independently informative resolution levels. Details are introduced next.

5.3 Blind Source Separation

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 through the structure of the index series, leads to *Independent Component Analysis* (ICA) [12, 17].

Combination these goals with the search for sparse signal representations suggests hybrid forms of *Sparse Component Analysis* (SCA) [39, 22, 48]. With SCA one attempts to combine the advantages delivered by sparsity and independence of signal representations, which transfer to better compression power and estimation in statistical minimax sense.

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.

Independent components can be efficiently computed by ad-hoc algorithms such as joint approximate diagonalization of eigenmatrices for real signals (*JadeR*) [13]. 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.

The *JadeR* algorithm is the ICA algorithm that we have applied in our experiments to deliver estimates for the separating or de-mixing matrix $B = A^{-1}$, and obtain the $Y = BX$, such that (under a perfect separation) $Y = BAS = S$. As a matter of fact, the solution holds approximately and up to permutation P and scaling D, i.e., $Y = DPS$. The decorrelation and rotation steps which have to be implemented will deliver a set of approximate m independent components.

5.4 Volatility Source Separation

We set the following system to represent a volatility process:

$$y_t = A_t x_t + \epsilon_t \tag{32}$$

where the observed returns are indicated by y_t, the mixing matrix A_t is to be estimated, together with the sources or latent variables x_t; the noise ϵ_t is superimposed to the system dynamics, with an i.i.d. $(0, \sigma_{\epsilon,t})$ distribution. We indicate with $v_t = \sigma^2_{\epsilon,t}$ the volatility process.

Ideally the volatility sources have a sparse representation, represented through the following system:

$$x_t = C_{jft} \Phi_{jft} + \eta_t \tag{33}$$

The compression and decorrelation properties of wavelet transforms can be supported by a more effective search for least dependent components through combined MRA and ICA steps. These steps can be set to work together in an hybrid method, as described below.

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.

One route is going through the iterations of the MP algorithm, by processing the observed returns with the WP and CP libraries, and by looking at:

$$Y_t \approx P_t \Phi_t + \xi_t = A_t C_t \Phi_t + \xi_t \tag{34}$$

where the noise is including an approximation error from the system equation and residual measurement effects ϵ_t.

We start by considering the detail signals obtained through WP and CP transforms, which refer to different degrees of resolution. 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.

We aim to optimize sparsity through the choice of ad hoc function dictionaries and optimze the performance of non-linear thresholding estimators. Furthermore, we search least dependent coordinates such that an almost diagonal covariance operator is achieved, helping the interpretation of latent volatility features.

5.5 Mixing Matrix Estimates

In Table 1 below we report 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 signals are passed through the ICA algorithm for the extraction of "m" possible sources which we set equal to the number of sensors.

We look at the results of this table so to extract from each detail level an approximate value indicating its contribution to the global signal features, independently from the other levels. The highest values computed suggest what are the dominant independent components 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., components belonging to other resolution levels, and only negligibly influenced by within-level factors. ICA selects the finest resolution levels of the WP Table, while for the CP Table it delivers a mix of components which are not concentrated at the finest resolutions.

The diagonalization achieved by ICA with WP delivers a resolution-wise ordered sequence of informative coefficients, with the highest scale delivering the most informative component, while for CP the order is not the same. Lower scale sequences are more Gaussian and stationary, due to aggregation effects, and thus it is likely that a certain loss of efficiency follows

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 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 6	0.0014	0.0046	**0.1748**	−0.0021	0.0036	−0.0052	0.002
level 2	−0.0089	−0.0023	−0.0031	**−0.1712**	0.0031	0.0057	−0.0006
level 5	0.006	0.0142	−0.0059	0.002	**0.1482**	−0.0053	0.0035
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**

Table 1: Weights of the estimated ICA mixing matrix A distributed across resolution levels for residual 5m series obtained in WP/CP tables.

from the fact of missing finer detail structure. Conversely, higher scale sequences, being locally more irregular series, might present better efficiency together with faster convergence rates.

The coefficient sequences obtained with WP and CP transforms have resolution-wise approximation power and thus perform more accurately, compared to the original series, with regard to the sample path from which the N data are observed. Processes for which a convergence rate N^{-1} is reached by pointwise continuous real valued functions are called irregular path processes [8] and the rate is called super-optimal. For stationary Gaussian processes with differentiable paths these rates are not achievable; similarly, lower scale sequences, compared to higher scale ones, reflect this aspect.

5.6 Realised-Integrated Volatility

Due to the fact that daily squared returns don't help too much in forecasting the latent volatility structure because of noise and of different dynamics, a measure has been suggested so to obtain a more accurate estimate of the volatility function with high frequency observations.

The *realised volatility* function is obtained by $\hat{\sigma}^2(t) = \sum_{i=1}^{T} r_i^2(t)$, and thus fulfills the scope of approximating the integral of the unobservable variable by averaging a certain number of 1m or 5m intraday values $r_i^2(t)$ [3, 4].

It holds [37] that $\hat{\sigma}_t^2 \to \sigma_t^2 = \int \sigma_s^2 ds$, i.e., the *integrated volatility* is approximated by the realised volatility obtained according to the *quadratic variation* (QV) principle derived from the following definition:

Definition 3: p^{th} variation
given X_t and the partition $T = \{t_0, t_1, \ldots, t_n\}$ of $[0, t]$, the p^{th} variation of X_t over T is:

$$V_t^{(p)}(T) = \sum_{k=1}^{n} \mid X_{t_k} - X_{t_{k-1}} \mid^p \tag{35}$$

If $\mid\mid T \mid\mid = max_{1 \leq k \leq n} \mid t_k - t_{k-1} \mid$ goes to 0, then $lim_{\mid\mid T \mid\mid \rightarrow 0} V_t^{(2)}(T) =< X >_t$, where the limit is the QV of X_t. This implies, in turn, that a convergence in probability applies, i.e., $\sum_{j=1}^{n} r_{n,t,j}^2 \rightarrow_{n \rightarrow \infty} \int_0^1 \sigma_{t+\tau}^2 d\tau$, where the cumulative squared high frequency returns are employed rather than the daily values, so to improve the volatility prediction power.

The QV principle and the realised-integrated volatility relation lead quite naturally [35] towards a non-parametric regression model as a *de facto* reference setting. With a normalized sampling index $t = \frac{i}{n}, i = 0, \ldots, n$, and with $y_{i/n} = n(x_{(i+1)/n} - x_{i/n})^2$ indicating noisy estimates of $n \int_{i/n}^{(i+1)/n} \sigma^2(x)_s ds$, from this latter expression one can recover $\sigma^2(x_t)$ and thus the whole procedure leads to a model like $y_t \approx \sigma^2(x_t) + \epsilon_t$.

Since both the regressors and the disturbances are not indipendent and identically distributed random variables, the volatility estimator can be found through families of non-parametric estimators where the search for more spatially homogeneous observational points require a re-mapping of the original values to some more regular domain, i.e., a ν-indexed grid such that $\hat{\sigma}_n^2(\bar{x}) = \sum_i^{\nu} K_{i,n}(x_\nu, \bar{x})$. Our next step is to identify efficient and feasible solutions for smoothing volatility.

5.7 Experiments

For pointwise volatility estimation a possibility is to use a parametric model such as a GARCH or SV, computed directly on the residuals obtained with the MP algorithm, after a certain number of steps [10, 11]. The focus is on detecting the time varying volatility features. We thus show parametric estimates from a model designed for the series of MP residues obtained after 100 and 500 iterations, i.e. by using 100 or 500 approximating structures.

We have adopted an MA(1)-GARCH(1,1) model, thus balancing serial correlation in the returns, and have used a Student's t as marginal distribution. The model should be written with a conditional mean equation (including an intercept and the MA(1) term in k) like:

$$y_t = k + \xi_t \sigma_t, \;\; with \;\; \sigma_t = \sqrt{h_t} \tag{36}$$

$$h_t = a_1 h_{t-1} + b_1 y_{t-1}^2 \tag{37}$$

where $y_t \mid \Psi_{t-1} \sim i.i.d.(0, h_t)$, $\xi_t \sim N(0, 1)$, and given the set of past information Ψ_{t-1}. By the prediction error decomposition, the log-likelihood function for a sample $y_1, \ldots, y_i$ is given by:

$$l_i(\Theta) = logL_T(\Theta) = \sum_{i=1}^{T} logp(y_i \mid \Psi_{i-1}) = -\frac{1}{2} \sum_{i=1}^{T} logh_i + \sum_{i=1}^{T} log \;\; g(\frac{x_i}{h_i^{1/2}}) \tag{38}$$

where g(.) is the Student's t distribution, an heavy tailed conditional distribution, given by $g(x) = c\frac{1}{(1+\frac{x^2}{v-2})^{\frac{v+1}{2}}}$, where $c = \frac{\Gamma(\frac{v+1}{2})}{(\pi(v-2))^{\frac{1}{2}}\Gamma(\frac{v}{2})}$.

Note that $x = y_i h_i^{-\frac{1}{2}}$ are the standardized residuals, and that the degrees of freedom are estimated together with the other parameters, say Θ, in the model. We then check diagnostic and statistical properties, reported in Table 2, where *t-stat* is calculated from the estimated standard deviations; *Deg. Fr.* refers to the return Student's t distribution; *LB* stands for Ljung-Box statistics for estimated squared standardized residuals; *MaxLik* is the estimated value of the likelihood function.

Parameters	(G)hf100w.res	(G)hf500w.res	(G)hf100c.res	(G)hf500c.res
MA(1)	0.30	0.19	0.33	0.21
t-stat	(33.87)	(18.53)	(30.08)	(17.7)
ARCH	0.012	0.014	0.076	0.078
t-stat	(9.95)	(8.51)	(9.6)	(5.96)
GARCH	0.973	0.98	0.699	0.20
t-stat	(341.75)	(387.8)	(25.89)	(1.895)
Deg.Fr.	2.88	5.29	2.84	5.16
LB	20.07	10.86	40.25	17.49
MaxLik	5327.87	7229.1	5176.29	6628.33

Table 2: GARCH (G) estimates for the residual series with 100 and 500 runs of MP with WP and CP tables indicated by *hf100/500w.res* and *hf100/500c.res*.

We observe that in general 500 iteration residues seem to suggest better models, as far as concerns likelihood estimated values and LB statistics. For CP and WP the estimates of the parameters change, particularly for the GARCH one, resulting for CP smaller in absolute terms and surprisingly not significant for the 500 iteration residues.

Some plots are now reported with regard to the quantiles with respect to the Student-t distributions of the residual sequences obtained by MP after 100 and 500 iterations. With 500 iterations the distributional properties suggested by the QQ-plots improve, for a better alignment with the ideal benchmark distributions.

Ideally, coherent structures should be removed and the algorithm should be stopped when dictionary noise is encountered. We observe second order statistical properties since features relate now to the volatility process characterizing the signal. When no structure is found this fact has to be interpreted as the evidence that only pure volatility aspects are left in the residual series, now clean of dependence, seasonal and non-stationary components.

Our results and diagnostic plots indicate that volatility is smoother when estimated from 500 MP iterations, compared to only 100. The possibilities are that it is either approximating the true volatility or that the MP algorithm is overfitting. For the CP case the same observation could be done, but due to the behavior of the GARCH parameter estimates, we might conclude that more than possible oversmoothing, MP has produced de-volatilization in the returns process.

Figure 4 is about the performance of the MP algorithm when examined through the

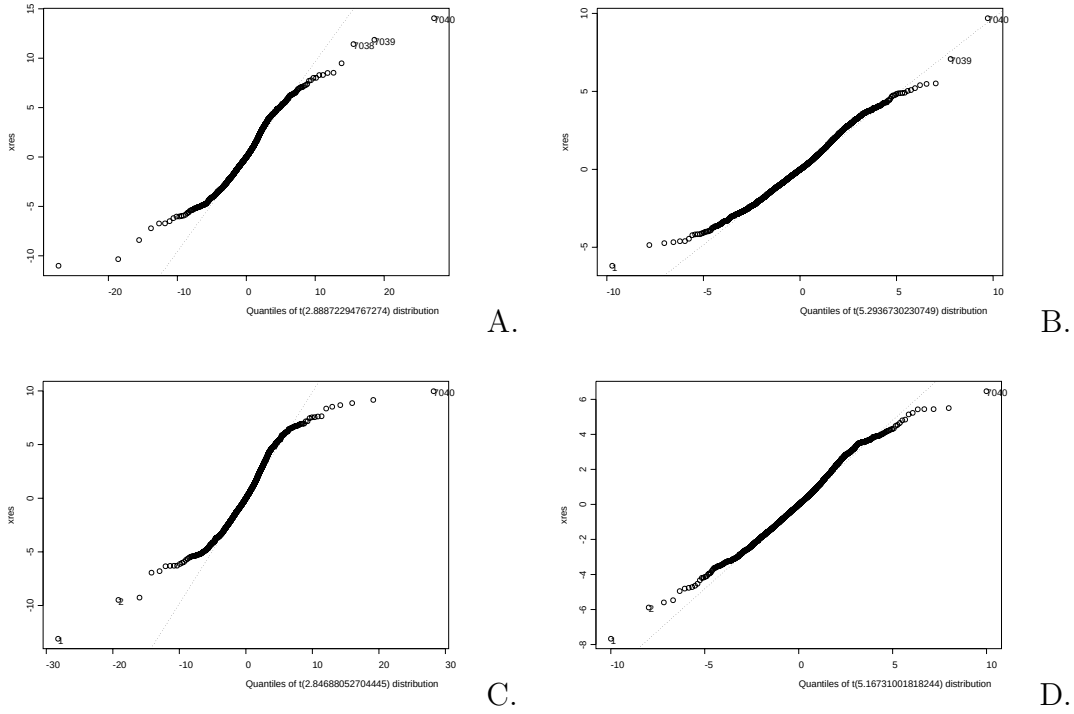


Figure 3: QQ-plots of standardized residuals vs quantiles of a Student's t, for 100 (A,B) and 500 (C,D) MP iterations run on respectively WP (left) and CP Tables. .

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. We 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 from the CP plots (C-D) that MP has a fast convergence; with both L_2 and L_1 norms, the first minimal turning point is at 100 structures, while the second one is at 200 structures. For the L_2 norm the successive decay is smooth, while for the L_1 norm is slightly steeper in approaching the new minimum at approximately 500.

From the WP plots (A-B) the limit of 100 structures is confirmed, but then convergence is lost. This confirms that by iterating 500 times the MP algorithm surely oversmooths the volatility function with WP approximating atoms, and instead of learning its structure it yields de-volatilization in the CP case.

We thus can conclude that the number of 100 iterations represents a fairly good benchmark iteration number of the MP algorithm for exploring the dynamics and learning the features of stock index return volatility, while at the same time preventing from possible numerical instability and overfitting.

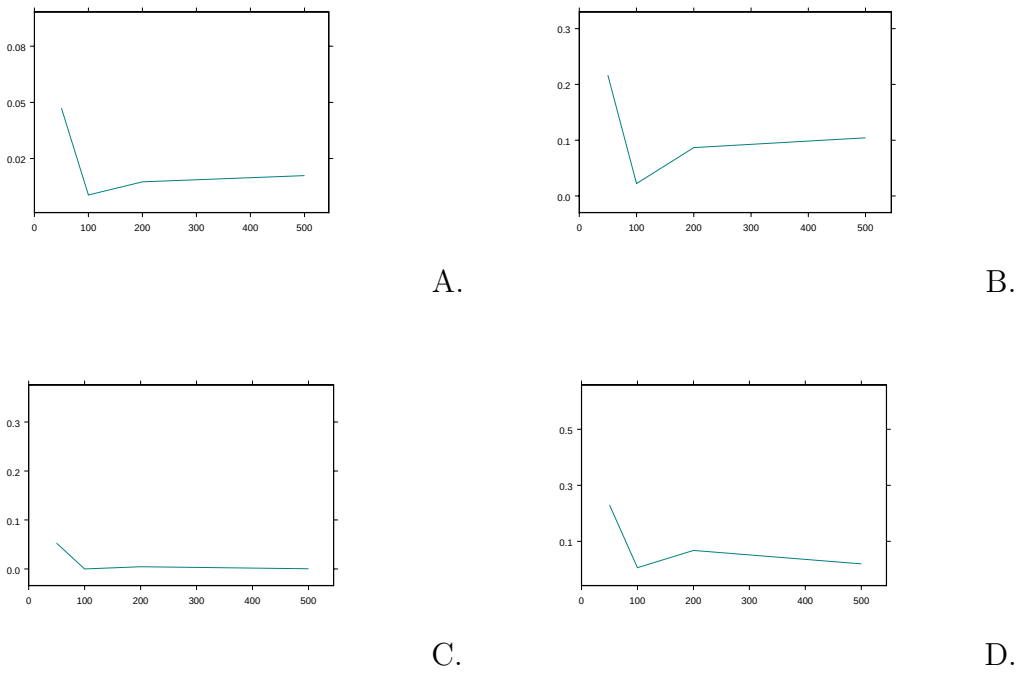


Figure 4: L_2 error vs number of approximating structures, for WP (A) and CP (C) and L_1 error vs L_0 norm for WP (B) and CP (D).

6 Conclusions

Financial time series are very complex structures and we show that in order to investigate their structure and detect volatility features it might be useful to adopt sparse representations and independent component decompositions, together with greedy approximation techniques. An example is offered by ICA applied to wavelet and cosine packets, and forming an hybrid form of SCA when combined with greedy feature learners. It appears from the experiments that the there is a clear improvement in the feature detection power of the latent volatility structure underlying a stock index return series.

The resolution selection operated by ICA, concentrated on high scale signals for WP function dictionaries, 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 scales.

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; 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.

The hybrid SCA method that we have designed for our application yields least dependent resolution levels which are used for calibrating the MP algorithm. The latter achieves a better detection power for the dependence structure in the return series due to the fact that the almost independent (or least dependent) coordinates along which the iterations are run,

allow the MP algorithm to be more orthogonalized, and thus more efficient in retrieving the most coherent atomic structures.

Acknowledgments

This work has been conducted mainly at CWI, Amsterdam (NL).

References

[1] 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" (C. Park, and W. Willinger, Eds.), Wiley, New York, (2000) 39-88.

[2] P. Abry, D. Veitch, and P. Flandrin, Long range dependence: revisiting aggregation with wavelets, *Journal of Time Series Analysis*, 19(3), (1998), 253-266.

[3] T. Andersen, T. Bollerslev, F.X. Diebold, and H. Ebens, The distribution of stock return volatility, *Journal of Financial Economics*, 61, (2001), 43-76.

[4] T. Andersen, T. Bollerslev, F.X. Diebold, and P. Labys, The distribution of exchange rate volatility, *Journal of American Statistical Association*, 96, (2001), 42-55.

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

[6] A. Benassi, Locally Self Similar Gaussian Processes, *in* "Wavelets and Statistics" (A. Antoniadis, and G. Oppenheim, Eds.), Springer-Verlag, New York, (1995), 43-54.

[7] T. Bollerslev, Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, 31, (1986), 307-327.

[8] D. Bosq, "Nonparametric Statistics for Stochastic Processes. Estimation and Prediction". Springer-Verlag, New York 1998.

[9] A. Bruce, and H.V. Gao, "S+Wavelets", StaSci Division, MathSoft Inc, Seattle 1994.

[10] E. Capobianco, Statistical Analysis of Financial Volatility by Wavelet Shrinkage, *Methodology and Computing in Applied Probability*, I(4), (1999), 423-443.

[11] E. Capobianco, Indipendent Multiresolution Component Analysis and Matching Pursuit, *Computational Statistics and Data Analysis*, 42(3), (2003) 385-402.

[12] J. Cardoso, Source separation using higher order moments, *Proceedings International Conference on Acoustic, Speech and Signal Processing*, (1989), 2109-2112.

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

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

[15] B. Cheng, and H. Tong, K-stationarity and wavelets, *Journal of Statistical Planning and Inference*, 68, (1998), 129-144.

[16] R.R. Coifman, Y. Meyer, and M.V. Wickerhauser, Wavelet Analysis and Signal Processing, *in* "Wavelets and their Applications" (M.B. Ruskai et. al., Eds), Jones and Bartlett, Boston, (1992), 153-178.

[17] P. Comon, Independent Component Analysis - a new concept?, *Signal Processing*, 36(3), (1994), 287-314.

[18] R. Dahlhaus, Fitting time series models to nonstationary processes, *The Annals of Statistics*, 25, (1997), 1-37.

[19] I. Daubechies, "Ten Lectures on wavelets", SIAM, Philadelphia 1992.

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

[21] G. Davis, S. Mallat, and M. Avellaneda, Greedy adaptive approximations, *Constructive Approximation*, 13(1), (1997), 57-98.

[22] D. Donoho, Sparse Components of Images and Optimal Atomic Decompositions, *J. Constructive Approximation*, 17, (2001), 353-382.

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

[24] D. Donoho, and I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *Journal of American Statistical Association*, 90, (1995), 1200-1224.

[25] D. Donoho, and I.M. Johnstone, Minimax Estimation via Wavelet Shrinkage, *The Annals of Statistics*, 26, (1998), 879-921.

[26] 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).

[27] R.J. Duffin and A.C. Schaeffer, A class of non-harmonic Fourier series, *Transactions of the American Mathenatical Society*, 72, (1952), 341-366.

[28] D. Donoho, M. Vetterli, R.A. DeVore, and I. Daubechies, Data Compression and Harmonic Analysis, *IEEE Transactions in Information Theory*, 44(6), (1998), 2435-2476.

[29] R.F. Engle, Autoregressive Conditional Heteroscedastic models with estimates of the variance of th UK inflation, *Econometrica*, 50, (1982), 987-1007.

[30] R.F. Engle, and G. Gonzales-Rivera, Semiparametric ARCH Models, *Journal of Business and Economic Statistics*, 9, (1991), 345-359.

[31] H.Y. Gao, Wavelet shrinkage estimates for heteroscedastic regression models, Tech. Rep., MathSoft Inc., Seattle, (1997).

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

[33] C.M. Hafner, Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models, *Journal of Statistical Planning and Inference*, 68, (1998), 247-269.

[34] T.J. Hastie, and R.J. Tibshirani, "Generalized Additive Models". Chapman & Hall, London, 1990.

[35] M. Hoffmann, L_p estimation of te diffusion coefficient, *Bernoulli*, 5(3), (1999), 447-481.

[36] I.M. Johnstone, and B.W. Silverman, Wavelet threshold estimators for data with correlated noise, *Journal Royal Statistical Socoety Ser. B.*, 59, (1997), 319-351.

[37] I. Karatzas, and E. Shreve, Brownian Motion and Stochastic Calculus, Springer-Verlag, New York, 1988.

[38] H. Krim, and J.C. Pesquet, On the statistics of Best Bases criteria, *in* "Wavelets and Statistics", (A. Antoniadis, and G. Oppenheim, Eds.), Springer-Verlag, New York, (1995), 193-207.

[39] M.S. Lewicki, and T.J. Sejnowski, Learning Overcomplete Representations, *Neural Computation*, 12(2), (2000), 337-365.

[40] S. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L_2(R)$, *Transactions of American Mathematical Society*, 315, (1989), 69-87.

[41] S. Mallat, "A wavelet tour of signal processing", Academic Press, 1999.

[42] S. Mallat, and Z. Zhang, Matching Pursuit with time frequency dictionaries, *IEEE Transactions in Signal Processing*, 41, (1993), 3397-3415.

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

[44] B.B. Mandelbrot, L. Calvet, and A. Fischer, The Multifractal model of asset returns, Disc. Pap. 1164, Cowles Foundation, Yale University, New Haven, (1997).

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

[46] T. Mikosch, and C. Starica, Limit theory for the sample autpcorrelations and extremes of a GARCH(1,1) process, *The Annals of Statistics*, 28(5), (2000), 1427-1451.

[47] M.H. Neumann and R. von Sachs, Wavelet Thresholding: beyond the Gaussian I.I.D. situation, *in* "Wavelets and Statistics", (A. Antoniadis and G. Oppenheim Eds.), Springer-Verlag, New York, (1995), 301-329.

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