Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Stochastic decomposition and approximation of stock index return volatility

E. Capobianco

Probability, Networks and Algorithms (PNA)

PNA-R0114 July 31, 2001

Report PNA-R0114
ISSN 1386-3711

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

Stochastic Decomposition and Approximation of Stock Index Return Volatility

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

ABSTRACT

This work is about applying wavelet-based approximation and estimation techniques to non-stationary financial time series for modeling stock index return volatility. The presence of various forms of dependence requires a careful analysis, particularly when dealing with very high frequencies and with periodic components. One important goal is achieving sparse signal decompositions, by the means of global and local function optimizers running through wavelet and cosine packet dictionaries, which are well suited for dealing with data of a complex nature. Another goal is obtaining a signal decomposition over statistically independent coordinates, so to let the algorithms learn in a more effective way the true structure characterizing volatility processes.

2000 Mathematics Subject Classification: 60H30, 62M10, 62G07.
Keywords and Phrases: Overcomplete Signal Representations; Greedy Approximations; Sparse and Independent Component Analysis; Volatility Estimation.

Note: This work was supported by ERCIM.

1. Introduction

Financial volatility models have time varying variance and covariance structures; this fact, while influencing modern theoretical, quantitative and computational finance based on simpler hypotheses, brings challenging problems from a statistical inference standpoint, due to the complexity of the non-linear, non-stationary and non-Gaussian properties involved.

One goal in this work is to represent volatility within the frame of so-called *Sparse Component Analysis* (SCA), proposed by modern signal processing and computational statistics techniques (Lewicki & Sejnowski, 2000; Donoho, 2001; Zibulevsky & Pearlmutter, 2001). This approach can represent an initial proposal for an innovative view of looking at volatility, and one goal of this paper is to suggest a model building strategy and verify its effectiveness in an application.

The main stylized facts about volatility models appear from many empirical studies (Mikosch & Starica, 2000), and among them the most important are heavy-tailed (leptokurtic) marginal return distributions; volatility clustering, and thus tail dependence; second-order dependence, visible in absolute and squared transformed returns, and pos-

sible long memory. They all represent features which are difficult to model, particularly with *ad hoc* statistical parametric or non-parametric procedures.
We deal with stock returns observed at very high frequencies from the Nikkei 225 composite index; the data were collected minute by minute and selected from an year (1990) among several available ones. Working with just one long realization of the underlying return process, no matter how precisely sampled, 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. We have approximately 35,000 observed values, covering intra-day market activity and excluding holidays and weekends; we then transform them in financial returns, by taking as usual the logarithms of ratios of consecutive time point prices. We also average the observed series, so to form a temporally aggregated signal, 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.
Dealing with non-stationarity is imposed by real circumstances; the observed series is very long, and thus subject to regime changes and external factors whose impact on the dynamics of returns and structure of volatility is undoubtedly relevant. Periods of stationarity alternate with others where some forms of system instability or time inhomogeneity affecting variables and parameters suggest that non-stationary behaviour is often expected and that it might be possibly approximated by looking at what we call (arbitrarily) a semi-stationary world, where periods showing smooth volatility patterns are combined with more random ones due to sudden and repeated fluctuations. This aspect strongly motivates the purpose of conducting data anaysis and doesn't prevent from pursuing the objectives of building more flexible model strategies.

2. Volatility Analysis

2.1 Classic Volatility Models

In the typical financial time series setting, simple models may involve complex structure. The heteroscedasticity behavior is due to the time-varying variance of the noise, whose squared value represents the volatility process. Volatility can be further specified as conditionally dependent on past squared returns and own lags, as for the class of *Generalized Autoregressive Conditionally Heteroscedastic* (GARCH) processes (Engle, 1982; Bollerslev, 1986; Taylor, 1986). When specified in a stochastic way, depending on an autoregressive structure together with a noise source which is independent from the disturbance term in the conditional mean equation, we refer to *Stochastic Volatility* (SV) processes (Ghysels et al., 1996).
It is possible that covariance non-stationarity is detected when investigating data series with these models; the presence of noise can also mask true features in the series, as in the case of latent (quasi-)periodic behavior, or emphasize the presence of features without identifying them appropriately as spurious components. As a consequence of observing non-stationarity, the underlying structure, possibly a mix of short and long range dependence (respectively SRD and LRD), becomes much harder to detect by diagnostic tools. Thus a reliable model cannot be designed without properly handling the observed data features (Andersen & Bollerslev, 1997a, 1997b).
The two indicated classes of volatility models are those which have gained diffuse acceptance; depending on the applications, GARCH or 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 $\sigma_t^2 =$

$\alpha_0 + \alpha_1 y_{t-1}^2 + \ldots + \alpha_p y_{t-p}^2$ (for an ARCH(p), in this case); thus, the information set Y_{t-1} is formed by past observations up to time *t-1*. In the other class the models are driven by parameters and include both observable and unobservable variables; the returns are distributed according to $y_t \mid h_t \sim N(0, exp(h_t))$, and the volatility is specified as $h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t$, $\eta_t \sim NID(0, \sigma_\eta^2)$. Usually one designs state-space models by transforming returns after a passage to the logarithms of their squared values. A formulation in latent components is obtained, as in (2) below, since the log-volatility h_t remains an unobserved process.
As a well-known remark, both classes of models can have ARMA structure; GARCH models can be written as non-gaussian linear ARMA in the squares of the observed returns, given a new process defined as $v_t = \sigma_t^2(\epsilon_t^2 - 1)$ (Bollerslev, 1986). This fact suggests that from the squared residuals we can get key diagnostic information about the volatility process, and thus calibrate our modelling strategies based on monitoring these residuals. As said, SV models allow volatility to depend on the unobserved components:

$$y_t = \epsilon_t exp(\frac{h_t}{2}); \quad h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t \tag{2.1}$$

with ϵ_t and η_t independent. Here h_t is a latent variable dependent on a random flow of new information entering the market at any time t. 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. Then, we can also model returns like linear ARMA:

$$log\ y_t^2 = h_t + log\ \epsilon_t^2; \quad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t \tag{2.2}$$

2.2 Flexible Models
In dealing with volatility more flexible models are available; they generally relax the distributional assumptions and functional forms involved. When few assumptions are retained, like ergodicity related to the mean reverting volatility (Fouque et al., 2000), and some form of dependence in the data, more or less strongly characterized, it may be sufficient to design a functional GARCH model with a volatility process generally described by $\sigma_t^2 = f(\sigma_{t-1}^2, \epsilon_{t-1}^2)$, or otherwise by a multiplicative model such as $\sigma_t^2 = \Pi_{i=1}^q f_i(\epsilon_{t-i})$, next to become additive via logarithmic transform.
With a non-parametric GARCH (Buhlmann & McNeil, 1999) one works with the functional model presented in a classic signal plus noise form, and the problem is that of recovering the unknown function combined with that of identifying its object, particularly when it represents stochastic volatility. Thus, it is also a deconvolution problem. One may end up with iteratively smoothing the unspecified function which characterizes the volatility dependence structure.
The model may be transformed in squared returns, then regressed on past original returns and estimated volatilities. Apart from choosing the optimal smoothing, and thus relying either on asymptotic considerations or adaptive plug-in procedures, the model is built with squared returns that can no longer exhibit long memory (Giraitis et al., 2000), thus lacking one possible key feature. While the LRD property is still not universally recognized as inherently natural to financial time series, its explicit exclusion is surely

hazardous. Furthermore, the model cannot work with stochastic versions of the volatility function, as reported in (Buhlmann & McNeil, 1999), for then consistency of the iterative smoothing procedure fails, given that it relies on the statistical independence between the information set with respect to the volatility and the information set based on the conditional mean noise process.
A multiplicative model (Hafner, 1998) is a simple and effective proposal, yielding an equation in logarithms for the transformed model and thus employing the use of backfitting estimation procedures (Hastie & Tibshirani, 1990). A certain robustness is shown despite the normal assumptions for the conditional mean noise process, while some more flexibility would be needed so to allow for structure in the volatility function, which instead of being simply a product (or sum, if logarithms are considered) of functions of past ARCH residuals, could also be dependent through forms of autoregression and be affected by independent noise.
As in the case of backfitting, which is inspired by the *projection pursuit regression* principle (Friedman & Stuetzle, 1981), the algorithmic procedure employed in the present paper is recursive, but operates through an orthonormal system, $\{\phi_{j_0 k}\}_{k\in Z} \cup \{\psi_{jk}\}_{j\geq j_0, k\in Z}$ of basis functions generated by dilations and translations of a scaling function ϕ_0 and a wavelet ψ. From such system a set of approximating wavelets may combine with thresholding rules so to deliver very effective and computationally fast nonlinear estimators, which for inhomogeneous function classes result superior to the linear ones employed by backfitting with some smoother. Then, the use of sparse representations for the signal, based on atomic decompositions of overcomplete dictionaries, makes the model structure much more flexible to handle and let near optimal minimax estimation go through.

2.3 Feature Detection
Figure 1 reports one-day 1m returns compared to the 5m aggregated values, together with their absolute and squared transforms. We note self-similar behavior and the typical shape conditioned to the different intensity of intra-day activity hours according to usual market technical phases. In defining *self-similarity*, and noting that (Mandelbrot et al., 1997) refer instead to the property of self-affinity, we follow (Benassi, 1995, §2, Def. 2.1) and have: *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* $R^{\gamma}_{\alpha,\bar{x}} f(x) = \frac{1}{\gamma^{\alpha}}[f(\bar{x}+\gamma x) - f(\bar{x})]\ \forall x \in R^d$ *and the following holds: given a Gaussian process* X *defined on a probability space* $(\Omega, \mathcal{F}, P)$, *it is a self-similar process of degree* α *if* $[R^{\gamma}_{\alpha,\bar{x}} X \stackrel{\mathrm{d}}{=} X], \forall \gamma \in R^{+} and \forall \bar{x} \in R^d$.

In Figure 2 we show the plots of the autocorrelation functions (ACF) computed over the entire sample size for the absolute and squared 1m and 5m raw returns.

One may note that:

- absolute and squared transformed series show an higher autocorrelation function, particularly in the 5m returns, with similarities and just differences in amplitude and time re-scaling;

- significant values are observed at distant lags, particularly with the absolute returns, suggesting the possible presence of long memory: the averaging effect of temporal aggregation is not changing the structure of dependence, where only initially we observe a fast exponential decay of the functions, followed by a slow decrease to zero, and thus suggests an hyperbolic behavior;

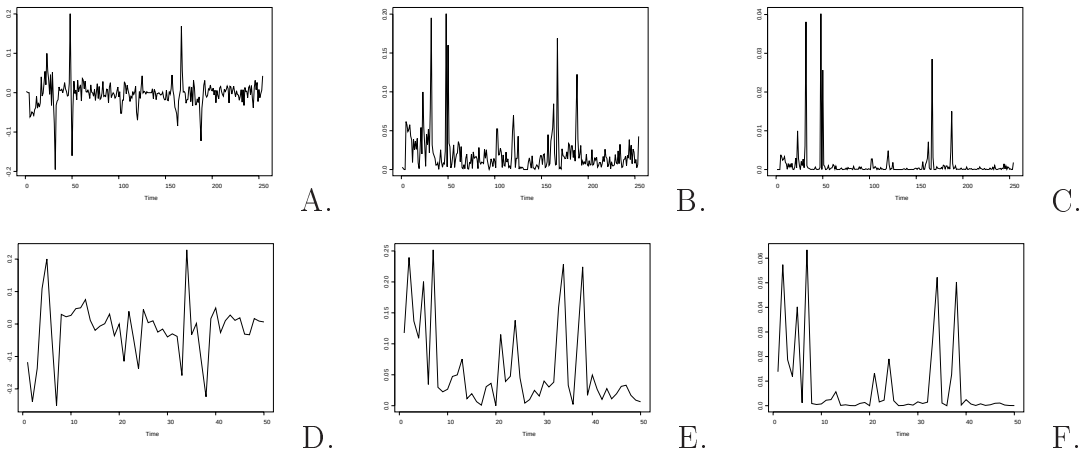


Figure 1: A) Raw 1m returns. B) Absolute 1m returns. C) Squared 1m returns. D) Raw 5m returns. E) Absolute 5m returns. F) Squared 5m returns.

- in the squared returns there is evidence of periodic or quasi-periodic behavior, due to a certain regularity of spikes, of a stochastic or deterministic nature.

3. Multiresolution Feature Learning

3.1 Wavelets

Given a *scaling function* or *father wavelet* ϕ, such that its dilates and translates constitute orthonormal bases for all the V_j subspaces obtained as scaled versions of the subspace V_0 to which ϕ belongs, and given a *mother wavelet* ψ together with the terms indicated with ψ_{jk} and generated by j-dilations and k-translations, such that $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$, one obtains differences among approximations computed at successively coarser resolution levels.

Thus, a so-called *Multiresolution Analysis* (MRA) 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 (Daubechies, 1992; Meyer, 1993; Hardle et al., 1998).

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

From Donoho (1996, $\S 4, pp.390$), *"when an orthogonal basis is an unconditional basis for a function space* F*, it means that there is an equivalent norm for the space,* $\|f\|_F$*, such that the ball* $\mathcal{F}(C) = \{: \|f\|_F \leq C\}$ *corresponds to a set of coefficient sequence*

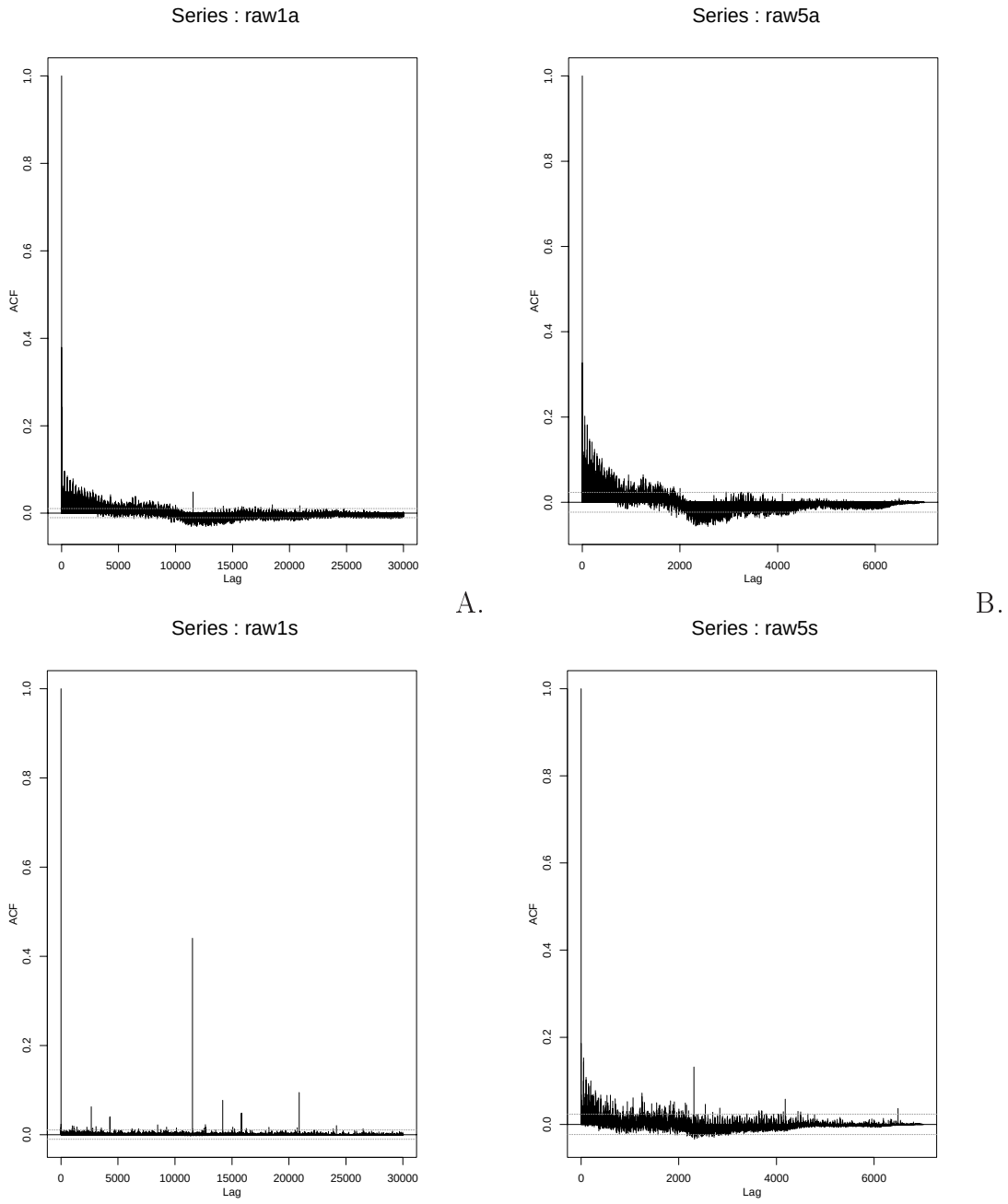


Figure 2: Absolute (indexed by a) and squared (indexed by s) raw 1m and 5m returns.

$\Theta(C) = \{\theta(f) : f \in \mathcal{F}(C)\}$ *which is solid and orthosymmetric"*, which means that if $\theta \in \Theta$ and $\mid \theta_i^{'} \mid \leq \mid \theta_i \mid, \forall i$, then $\theta^{'} \in \Theta$. As a consequence, an unconditional basis diagonalizes a functional class and retains optimal sparsity (see also Donoho, 1993).
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:

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

where $\phi_{j0,k}$ is associated with the corresponding coarse resolution coefficients $c_{j0,k}$ and $d_{j,k}$ are the detail coefficients, i.e. $c_{j,k} = \int f(x)\phi_{j,k}(x)dx$ and $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$. In short, the first term of the right hand side of (3) is the projection of f onto the coarse approximating space V_{j0} while the second term represents the cumulated details.

De-noising In the wavelet-based representations of signals sparsity inspires strategies that eliminate redundant information, not distinguishable from noise; this can be done in the wavelet coefficients domain, given the relation between true and empirical coefficients, $\tilde{d}_{jk} = d_{jk} + \epsilon_t$. The *wavelet shrinkage principle* (Donoho & Johnstone, 1994, 1995, 1998) applies a thresholding strategy which yields de-noising of the observed data; it operates by shrinking wavelets coefficients toward 0 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.

The following well-known algorithm is usually implemented:

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

The *shrinkage rule* and the *threshold value* are selected among several possible choices, and given the noisy nature of observed financial time series, an *adaptive procedure* might be preferred. The *soft shrinkage* rule selected is $\delta_s(\tilde{d}_{jk}, \lambda) = sgn(\tilde{d}_{jk})(\mid \tilde{d}_{jk} \mid -\lambda)_+$, when $\mid \tilde{d}_{jk} \mid > \lambda$, or otherwise $\delta_s(\tilde{d}_{jk}, \lambda) = 0$. It thus keeps or shrinks values, compared to the keep-or-kill solution offered by the *hard rule*, where $\delta_h(\tilde{d}_{jk}, \lambda) = \tilde{d}_{jk} I(\mid \tilde{d}_{jk} \mid \geq \lambda)$.

Wavelets and Finance Inhomogeneous functional classes characterization, diagonalization power and sparsity yield, together with the MRA property, a powerful approximation instrument. In representing a function belonging to a general space, an increased localization power yields advantages in terms of spatial adaptivity, which might be very useful

for handling financial time series, where the relevance of second order statistical information related to covariance function estimation, suggests that its diagonalization property may be very useful, especially in multivariate contexts or when dealing with statistical efficiency of parameter estimators.
Following (Capobianco, 1999a) we have compared the performance of some learning procedures running on wavelet- and cosine-based dictionaries; with such dictionaries complete or overcomplete signal representations may be obtained, meaning that redundant wavelet coefficients are produced or not, compared to the size of the original signal. The consequencies are for the bases that may be possibly formed, resulting unique, in the complete case, or multiple ones.
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 the temporal coordinate. Thus, from a structure with LRD one finds SRD or a decreased dependence, when looking at wavelet coefficients at each scale 2^j or resolution level j. 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 (see Johnstone and Silverman, 1997).
An important aspect is non-stationarity, the almost natural condition of financial time series, especially when measured at very high frequencies. A long time series is, as a matter of fact, subject to shifts and regimes from various factors, related to the return specific dynamics and external influence of many factors and variables. At high frequencies, even at the intra-daily period of observation, some trading dynamics suggest that non-stationary behavior can come in the form of hidden periodic or quasi-periodic behavior. 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 (Cheng & Tong, 1998) $X(t)$ is a stationary process if and only if it is stationary at all levels of resolutions (Prop. 3.4), while stationarity at the kth level of resolution (or k-stationarity) occurs (Def. 3.1) if $\forall n \geq 1, t_1, \ldots, t_n \in T$ and $l \in Z$, it holds that $[X(t_1 + 2^{-k}l), \ldots, X(t_n + 2^{-k})] \stackrel{d}{=} [X(t_1, \ldots, X(t_n)]$. Given a representation through a continuous scaling function family ϕ_{kl} such that, $X(t) = \sum_l \eta_{kl}\phi_{kl}(t)$ they state that the condition for stationarity at level k (Prop.3.1) depends on $\eta_k(.)$ being stationary, i.e. $\eta_k(.) \stackrel{d}{=} \eta_k(. + l), \forall l \in Z$.
From (Abry et al., 2000), given a stochastic process X featuring LRD or strong dependence structure, the following holds, given $E[d_x(j,k)] = 0$, for the variance:

$$E[d_x(j,k)^2] = \int , _x(v)2^j \mid \Psi_0(2^j v) \mid^2 dv \tag{3.2}$$

representing a measure of the power spectrum $, _x(.)$ at frequencies $v_j = 2^{-j}v_0$, with Ψ_0 the Fourier Transform of ψ_0. Given $, _x(v) \sim c_f \mid v \mid^{-\alpha}$ and $v \to 0$, then $E[d_x(j,k)^2] \sim 2^{j\alpha}c_f C(\alpha, \psi_0)$, for $j \to \infty$, and $C(\alpha, \psi_0) = \int \mid v \mid^{-\alpha} \mid \Psi_0(v) \mid^2 dv$, for $\alpha \in (0,1)$.
The covariance function of the wavelet coefficients is instead controlled by N, the number of vanishing moments, which when are present in sufficiently high number lead to high compression power for the fact that the finest coefficients are negligible for smooth part of the function; in particular, the decay is much faster in the wavelet expansion coefficients domain than in that originated by LRD processes.

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

$$E[d_x(j,k)d_x(j,k') \approx \mid k-k' \mid^{\alpha-1-2N}, for \mid k-k' \mid \to \infty \tag{3.3}$$

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 $(j,k) \neq (j,k')$. The immediate consequences are that an average measure with small variance can be formed, $\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} \mid d_x(j,k) \mid^2$, which results an unbiased estimate of the variance of the detail signals, for each j. Depending on the scaling factor $\frac{1}{n_j}$, related to the dimension of the level j, this average value is also asymptotically efficient, together with preserving the power law dependence at j in the variance computed for each scale.

Temporal Aggregation There is a clear relation between time aggregation and resolutions, which establishes that with lower scales (higher resolution indexes and coarser details) a more extended time horizon is retained between individual observations, with increased smoothing power, while with higher scales and lower resolutions (finer details) the temporal distance between sample values is reduced and thus more high frequency information is accounted for.
By defining a T-aggregated process as follows, $X^T(n) = \frac{1}{T}\int_{(n-1)T}^{nT} X(t)dt$, this process can be investigated with wavelets, thus giving a relation between time aggregation and multiresolution analysis (Abry et al., 1998). From the detail signals, one can form the variance as follows: $var(d_x(j,k)) \approx 2^{j\alpha}$, for $j \to \infty$. For a LRD process, the autocorrelation function assumes the form $C_x(\tau) \sim \tau^{-\beta}$, for $\tau \to \infty$ and $\beta \in (0,1)$. The case $\beta \in (1,2)$ suggests that SRD is present, even if with power law structure, thus remaining $var(d_x(j,k)) \sim 2^{j(1-\beta)}$, for $j \to \infty$ and each k. Looking at the autocovariance function, given j, the law is now $\mid k-k' \mid^{-\beta-2N}$, thus always SRD due to the decorrelation effect of wavelets enabled by the control of non-stationarity and dependence offered through the N parameter.
This last fact has consequencies when modeling the power law of wavelet coefficients variances and covariances in relation with the number of vanishing moments of the wavelet family, suggesting that the choice of an appropriate number N is the key variable for dealing with LRD when a wavelet transform is adopted. The variance of the designed estimators benefits from the flexibility allowed by the wavelet system, particularly with regard to an improved asymptotic efficiency, because less dependence is left after the transform, even if a more Gaussian-wise behavior at coarser details comes with the price of a loss of high frequency information.

Overcomplete Representations Function dictionaries are collections of parameterized waveforms or atoms (Chen et al., 2001); they are available for 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 serve to represent remaining dictionary structures, thus originating a non-unique signal decomposition.

An example of overcomplete representations is offered by wavelets packets, which represent an extension of the wavelet transform and allows for better adaptation due to an oscillation index f related to a periodic behaviour in the series which delivers a richer combination of functions. Given the admissibility condition $\int_{-\infty}^{+\infty} W_0(t)dt = 1$, $\forall(j,k) \in Z^2$ we have, following (Krim & Pesquet, 1995):

$$2^{-\frac{1}{2}} W_{2f}(\frac{t}{2} - k) = \sum_{i=-\infty}^{\infty} h_{i-2k} W_f(t-i) \tag{3.4}$$

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

$$2^{-\frac{1}{2}} W_{2f+1}(\frac{t}{2} - k) = \sum_{n=-\infty}^{\infty} g_{n-2k} W_f(t-n) \tag{3.5}$$

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) through $\{2^{-\frac{j}{2}} W_f(2^{-j}t - k), k \in Z, (j,f) \mid I_{j,f} \in P\}$. 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 *Wavelet Packet Transform* (WPT).
With a *Cosine Packet Transform* (CPT) we have good bases as far as concerns compression power, as shown by (Donoho et al., 1998), thus getting sparsity of representations. In (Mallat et al., 1998) it is shown that CP are optimal bases for dealing with non-stationary processes with time-varying covariance operators. The building blocks in CP are localized cosine functions, i.e. localized in time and forming smooth basis functions. They are almost eigenvectors of locally stationary processes (the latter defined in Dahlhaus, 1997; Neumann & von Sachs, 1995) and thus constitute almost diagonal operators used to approximate the covariance function. It is common to represent a signal from WP dictionaries as $f(t) = \sum_{jfk} w_{j,f,k} W_{j,f,k}(t)$ and of CP ones as $f(t) = \sum_{jfk} c_{j,f,k} C_{j,f,k}(t)$.
The CPT has an advantage over the classic *Discrete Cosine Transform* (DCT); the latter defines an orthogonal transformation and thus maps a signal from the time to the frequency domain, but it is not localized in time and thus is not able to adapt well to non-stationary signals. Depending on the *taper* functions we select, the cosine packets decay to zero within the interval where they are defined and in general determine functions adapted to overcome the limitations of DCT. A *DCT-II* transform is defined as:

$$g_k = \sqrt{\frac{2}{n}} s_k \sum_{i=0}^{n-1} f_{i+1} cos(\frac{(2i+1)k\pi}{2n}) \tag{3.6}$$

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. The within-block coefficients of the WP and CP formulations describe their contribution in representing the signal features under a varying oscillation index. The WP Table (Figure

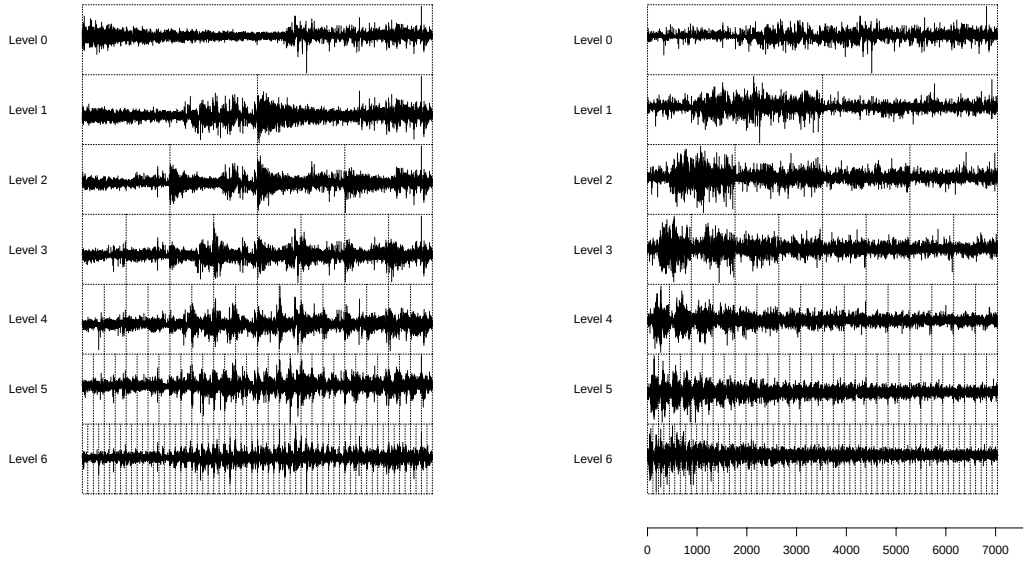


Figure 3: CP table (A) and WP table (B) with signal segmentation level-by-level.

3, B) presents crystals, i.e. sets of coefficients, stored in sequency order, according to increasing oscillation index. The CP Table (Figure 3, A) presents instead blocks ordered by time and the coefficients within the blocks are ordered by frequency. Thus, in the WP Table 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 (the oscillation index f goes from 0 to $2^J - 1$, going rightwise). For the CP Table instead, the high frequency part of the signal is now expected on the left side, while the low frequency behavior appears from the right side.

3.2 The Matching Pursuit Learning Algorithm

The design of optimal algorithms is strictly dependent on the adoption of adaptive signal approximation techniques, built on sparse representations. Sparsity refers to the possibility of considering only few elements of a dictionary of approximating functions selected among a redundant set. The MP algorithm (Mallat & Zhang, 1993) is a good example, and it has been successfully implemented in many studies for its simple structure and effectiveness. A signal is decomposed as a sum of atomic waveforms, taken from families such as Gabor functions, Gaussians, wavelets, wavelet and cosine packets, among others. We focus on the WP and CP tables, whose signal representations are given by:

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

and

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

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

- the approximating kernels are flexible with regard to the type of functions used, i.e. localized cosine functions and variably oscillating wavelets;

- the mixtures of functions employed work in space/time and scale/frequency dimensions, thus yielding better spatial adaptivity and localization power;
- a priori or signal-dependent knowledge may be accounted for, by selecting indexed functions or by reducing the problem dimension through the use of a restricted sub-set of functions in the analysis.

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

$$f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + res_n(t) \tag{3.7}$$

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 nth MP coefficient h_n (or $\mu_{\gamma_n,n}$) and atom H_{γ_n} the computation of the updated nth residual is given by:

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

5. repeat the procedure from step 2, 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, see Mallat & Zhang, 1993). 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 , $_0 \in$, .
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 (Davis et al., 1997) *coherent structures* compared to the *dictionary noise* components. In our application another 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.

3.3 The Best Basis Algorithm

The *Best Orthogonal Basis* (BOB) algorithm (Coifman & Wickerhauser, 1992) is here employed as an alternative to the MP optimization method, with the goal of minimizing an additive cost function computed within a library of orthonormal basis representations generated by the WP and CP transforms and through the correspondent expansion coefficients. The procedure adaptively picks the best orthogonal basis among those which can be formed as sub-collections of WP or CP dictionaries. The BB algorithm thus represents a global optimizer which computes the transform by searching for the minimum of a cost function $E(C) = \sum_{j,f} E(w_{j,f})$ in O(LN) operations, with $L = log_2 N$ the number of levels of the binary tree and N is the signal length (this compared to the O(MLN) cost of the MP, with M packets selected).

In particular, the BOB steps find a minimum entropy transform from the dictionary at hand, since the above objective function corresponds to $min\ [entr\ f(B)] \mid B \in ,$, where B is an orthobasis in the selected dictionary , and $f(B)$ are a vector of coefficients in the same basis. In terms of the entropy, commonly used in statistics for estimation and compression problems, the cost function holds as $E^{ent}_{j,f} = \sum_k \hat{w}^2_{j,f,k}\ log\ \hat{w}^2_{j,f,k}$, for $\hat{w}_{j,f,k} = w_{j,f,k} \times (||\ w_{0,0}\ ||_2)^{-1}$. The algorithm is known to deliver near-optimal sparsity representations, but not in the presence of non-orthogonal contexts.

In Figure 4 we report about the top-100 largest coefficients approximation with the BB and the MP algorithms after running on WP and CP dictionaries. We show the BB on the WP table in (A), and on the CP table in (B), while for the MP algorithm we report respectively in (C) and in (D). The plots suggest that BB doesn't work optimally for the non-stationary signal, while MP works more efficiently; this is due to its greedy nature, and it results more effective for a better ability to capture the local features, both in time and in frequency. The MP scheme exploits the correlation power inherent to the collection of waveforms available through the WP and CP dictionaries, and it does so throughout more scales and by extending the basis which represents the signal.

4. Independent and Sparse Component Analysis

4.1 General Aspects

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 data, leads to *Independent Component Analysis* or ICA (Cardoso, 1989; Comon, 1994; Jutten & Herault, 1991). The combination of these goals with that of searching for sparse signal representations suggests hybrid forms of SCA. With SCA one attempts to combine the advantages delivered by sparsity of signal representation, which transfer to better compression power and estimation in statistical minimax sense.

We present the results obtained with ICA, whose role has gained huge relevance to applications in many fields, particularly signal processing and neural networks. The independent components can be efficiently computed by ad-hoc algorithms such as *JadeR* (Cardoso & 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. Therefore, ICA is a latent variable statistical model where linear or non-linear transforms of non-Gaussian and independent variables deliver the observed data.

By assuming that the sensor outputs are indicated by $x_i, i = 1, \ldots, n$ and represent a combination of independent, non-Gaussian and unknown sources $s_i, i = 1, \ldots, m$, a non-

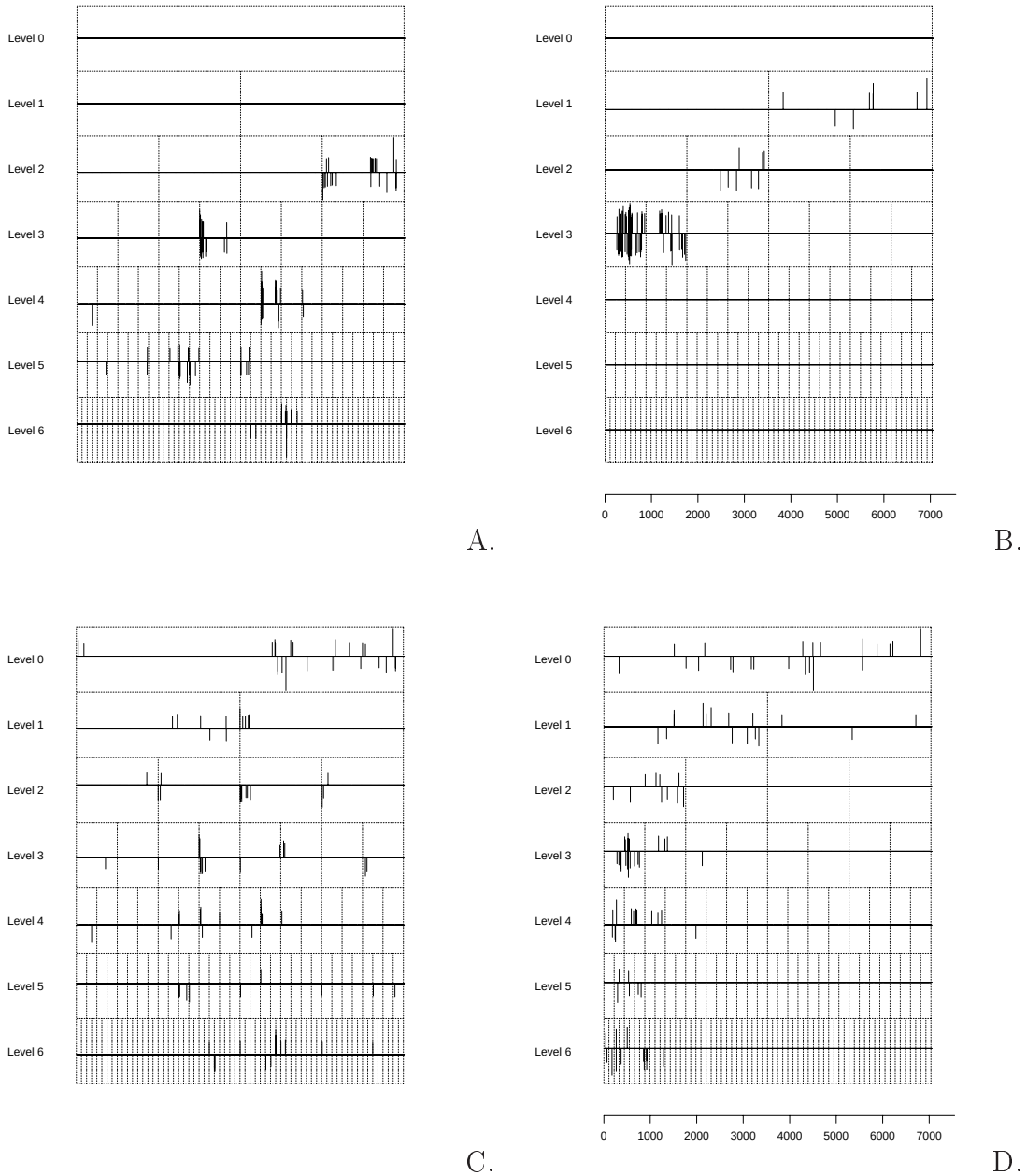


Figure 4: Signal approximation with the 100 largest coefficients for BB run on CP (A) and WP (B), and for MP run on CP (C) and WP (D).

linear system $Y = f(X)$ could be approximated by a linear one AS, where $X = AS$. Instead of computing $f(X)$ one may now work for estimating the sources S together with the $m \times m$ mixing matrix A, where usually $m << n$, with n the number of sensor signals, but with $m = n$ holding in many cases too.
The *Joint Approximate Diagonalization of Eigenmatrices for Real signals* (i.e. JadeR) is the algorithm that we have chosen to implement ICA; it delivers an estimate for the separating or de-mixing matrix B, obtained from $Y = BX$, such that when $B = A^{-1}$ a perfect separation would be obtained. This in general cannot happen, being just an ideal setting, and thus solutions hold approximately up to permutation and scaling. De-correlation and rotation steps are implemented so to deal with these aspects, and a set of approximately m independent components is obtained.

4.2 Relationships with Factor Models
Statistically independent components may offer a possible interpretation of the main driving forces behind financial time series, in line with other decomposition techniques such as structural time series analysis or factor models. Many factor models have been suggested, like that proposed by (Ray & Tsay, 2000) who question if LRD is a common factor among stocks which can be described by long memory SV models (Breidt et al., 1998). The basic frame starts by considering one of such models, where $y_t = \sigma_t \epsilon_t$ and $\sigma_t = \bar{\sigma} exp(\frac{v_t}{2})$, with $\bar{\sigma}$ a positive constant, ϵ_t an iid(0,1) process and v_t a fractionally integrated I(d) process with $d \in (0, \frac{1}{2})$, i.e. $(1-B)^d v_t = \eta_t$. With v_t Gaussian, σ_t^2 is the volatility measure. A standard model transformation applies, just by taking logarithms of squared returns, $z_t = \mu + v_t + \eta_t$, thus being a Gaussian long memory signal plus non-Gaussian noise.
A generalization of this model is to allow a k-variate common component model, with the simple passage to the vector equation, $z_t = H v_t + \xi_t$, the v_t process now representing a reduced (compared to k) r-dimensional vector addressing the common LRD structure, and thus H representing a $(k \times r)$ matrix of rank r. A matrix $\tilde{H}$ can be found such that its transpose multiplied by H gives a zero matrix, leaving the common factor model reduced to $z_t = \tilde{H} \xi_t$, i.e. a non-Gaussian but SRD process, given that the long memory contribution from the v_t process has been eliminated.
How to find that $\tilde{H}$, or otherwise that linear combination of transformed returns with SRD structure, is the important methodological question now. The authors proposed a canonical correlation method and a related test which allows to identify the case when SRD is achieved. We have elaborated the model dimension reduction idea in a different setting, that of univariate models, and for different scopes, like the search for independent coordinates so to diagonalize the covariance operator. Nevertheless, the same simple idea yields nice results when applied.
Dictionaries which are overcomplete deliver non-unique signal decompositions; when instead a basis may be selected, the dictionary will result complete. In our applications the hybrid method we have designed requires that the least dependent resolution levels are to be selected by ICA and used for calibrating the MP algorithm, thus achieving a better detection power for the dependence structure in the series. More independent coordinates along which to apply the algorithmic steps allow the MP to be more orthogonalized and thus work more efficiently in retrieving the coherent structures; the algorithm learns more effectively, working progressively toward obtaining a final residue whose absolute and squared transforms might reveal only pure volatility features.

4.3 A Volatility Frame

To summarize, we have the following system to represent a volatility process:

$$y_t = A_t x_t + \epsilon_t \tag{4.1}$$

where the observed returns are indicated by y_t, the mixing matrix A_t is to estimated, together with the sources or latent variables x_t; the noise ϵ_t the 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:

$$x_t = C_{jft}\Phi_{jft} + \eta_t \tag{4.2}$$

We lose the typical autoregressive form of dependence, but leave the structure to be non-parametrically investigated by selected dictionaries of functions, wavelet packets and localized cosines. We also maintain in (10) the underlying hypothesis that a mixture basic law of information arrivals is governing the market dynamics.

Since the sources are unobservable, estimating them and the mixing matrix is quite complicated; we can either build an optimization system with a somewhat regularized objective function through some smoothness priors, so to estimate the parameters involved, or we can proceed more recursively in the mean square sense, through iterations of the MP processing the observed returns with the WP and CP libraries, and looking at $y_t \approx P_t\Phi_t + \xi_t = A_t C_t \Phi_t + \xi_t$, where the noise is including system η_t and residual measurement effects ϵ_t. The MP algorithm delivers sparse P_t by the means of a denoising step related to the orthonormal wavelet system selected, but remains unable to disentangle the components of the operator. It will be, according to our strategy, up to an ICA step, to deal with this aspect.

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 actions. The strategy here considered is the following. We start from considering the detail signals obtained through WP and CP transforms. The series of projected scaled signals refer to different degrees of resolution and reflect specific information obtained by the transforms while switching between resolution levels. Then, we combine an ICA step with the MP algorithm operating on WP and CP tables; through such a joint search for sparsity and statistical independence we are basically adopting an hybrid SCA solution, since we aim to optimize sparsity through the choice of ad hoc function dictionaries, like localized cosines and orthonormal wavelet bases, and because we also adopt non-linear thresholding estimators. Furthermore, we want to operate through least dependent coordinates such that an almost diagonal covariance operator is achieved, helping the interpretation of latent volatility features.

This frame allows us to link also to the factor model presented before. With univariate series, the common persistence is that observed not from many stock sequences but from details obtained from the common initial signal, the stock index series. Since they all inherit persistence and long memory, the decorrelation induced by the wavelet transform first, and then by ICA, permits to associate $y_t = A_t C_t \Phi_t + \xi_t$ to $z_t = \tilde{H}\xi_t$, where the product of A_t and C_t produces the same effect of delivering SRD residuals through transformed linear combination of initially observed returns.

4.4 Interpreting the Model

One result that was obtained is the following, as we can see in the reported experiments: ICA applied to WP detail signals achieves a better compromise between obtaining a sparse representation and a set of least dependent components. By looking at results in the next section, the selection of high scale signals eliminates redundant information by keeping highly localized time resolution power without simultaneously losing too much frequency resolution. The denoising step too, here applied, permits to improve the S/N ratio considerably, and thus deliver a sparser signal representation.

The independent components, as said, not necessarily exist, particularly with non-stationary and dependent signals; one must turn to other devices, like combining ICA with wavelet-based signal decomposition and de-noising, so to form an hybrid SCA. In this way one finds that for non-Gaussian data wavelet-represented signals result the least dependent components selected, and a formal representation is given by $x(j,t) \approx \sum_j \theta_j(t)\Phi_j(t)$, i.e. sequences obtained by a sequential application of wavelet and cosine transforms, through the operator Φ_j, and ICA, applied on θ_j.

When ICA is applied to a CP library, it doesn't really concur to build a sparse representation, since the CP coordinates are already naturally endowed with that property; from one aspect it depends on the time domain segmentation operated according to the degree on discontinuity revealed by the data. Thus, for a certain time interval, the size of the local cosine windows might correspond well to that representing an approximate stationary behavior for the process at hand.

From (Mallat et al., 1998) we know that local cosine vectors might be approximate eigenvectors of the covariance operators and that an orthogonal basis of them yields a sparse matrix with fast off-diagonal elements decay when a locally stationary process is observed. This sparse matrix should be estimated and ideally might be assumed to be a band or near diagonal matrix; one solution is BOB, but we have already seen that for our time series is sub-optimal compared to the greedy MP.

5. Non-parametric Estimation

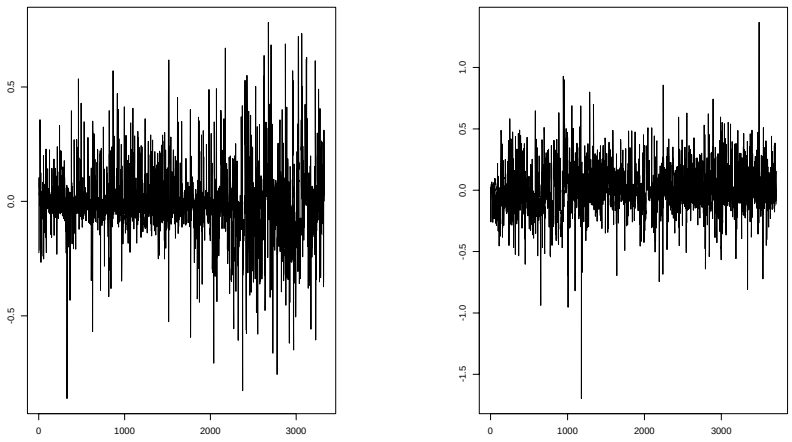
5.1 Semi-Stationary Covariance Processes

With the expression "semi-stationary" we are not going to define properties that characterize a certain class of processes, like with locally stationary processes, but instead address the fact that a lack of stationarity occurs in periods or timelengths which are different depending on the presence of regime shifts or shocks or other independent factors, then reverting to a more stationary behaviour.

We have run experiments so to test the MP approximation power, and we let the algorithm work with 50, 100, 200 and 500 atoms from the selected WP and CP function dictionaries, so to verify whether the residual structure might be better interpreted and might suggest how efficient is learning when we change some conditions. De-noising the tables via thresholding is another useful way of testing how the MP performance is influenced by the presence of noisy wavelet crystals. For thresholding under non-standard statistical conditions, we have found relevant contributions from the work of (Gao, 1997; Johnstone, 1999; von Sachs & MacGibbon, 2000). We have adopted the following steps:

Step 1

We apply a wavelet/cosine packet table segmentation, thus splitting the initial sample into segments, ideally thought to be more stationary than the whole series. The chosen procedure delivers certain computational advantage and an improved local fit power to



A. B.

Figure 5: First sub-series, n=1-3328 (A) and second sub-series, n=3329-7040 (B).

be used for estimating the variance by looking more specifically at the data dynamics belonging to less non-stationary segments, which may correspond to separate market phases. The simplest segmentation is applied by reflecting the sample splitting rule which restricts to sample sizes divisible by 2^J in the wavelet/cosine packets. We keep it at its minimum level, by just considering two return sub-samples (see Figure 5).

Step 2

The following thresholding algorithm is adopted:

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

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

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

and through the following functional:

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

can find a nonlinear function estimator like:

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

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

Step 3

Then *we apply the MP algorithm to the sub-tables*, i.e. to the sample segments previously computed. We run MP with increasing approximation power and in both the original tables and their waveshrunken versions.

Step 4

We check how the approximation power of the MP algorithm is affected by the noise, by looking at the usual diagnostic ACF plots for the absolute and the squared residuals (Figure 6). 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 conditional variance, or 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 SRD, LRD, seasonal and non-stationary components.

The ACFs computed over the de-noised residuals indicate that the noise, somehow spuriously, contributes to the structure shown by the WP/CP tables. In summary, the noise seems to hide periodic components and its removal allows for a better detection of them, and this suggests that non-stationarity is very likely responsible for the presence of spurious features.

5.2 Time and Frequency Resolution Pursuit

Together with the risk of finding spurious components for the non-stationary nature on the data, the masking effects of noise has been indicated as a further difficulty in dealing with high frequency financial time series. These factors should be considered combined with possible overfitting effects when the MP optimization procedure is run; the algorithm could learn too much and adapt even to non-features. De-noising through the SURE-Shrink estimator alone may not be sufficient; an important aspect concerns the structure of the algorithm itself, and its pursuit activity throughout the resolution levels. We want to focus on how to optimally exploit the information coming from the most informative part of the signal, with regard to both its time and frequency content. Thus, we investigate the performance of the MP algorithm when is applied on a restricted and *ad hoc* selected range of resolution levels.

In Table 1 below we report the two estimated mixing matrices A, where the observed

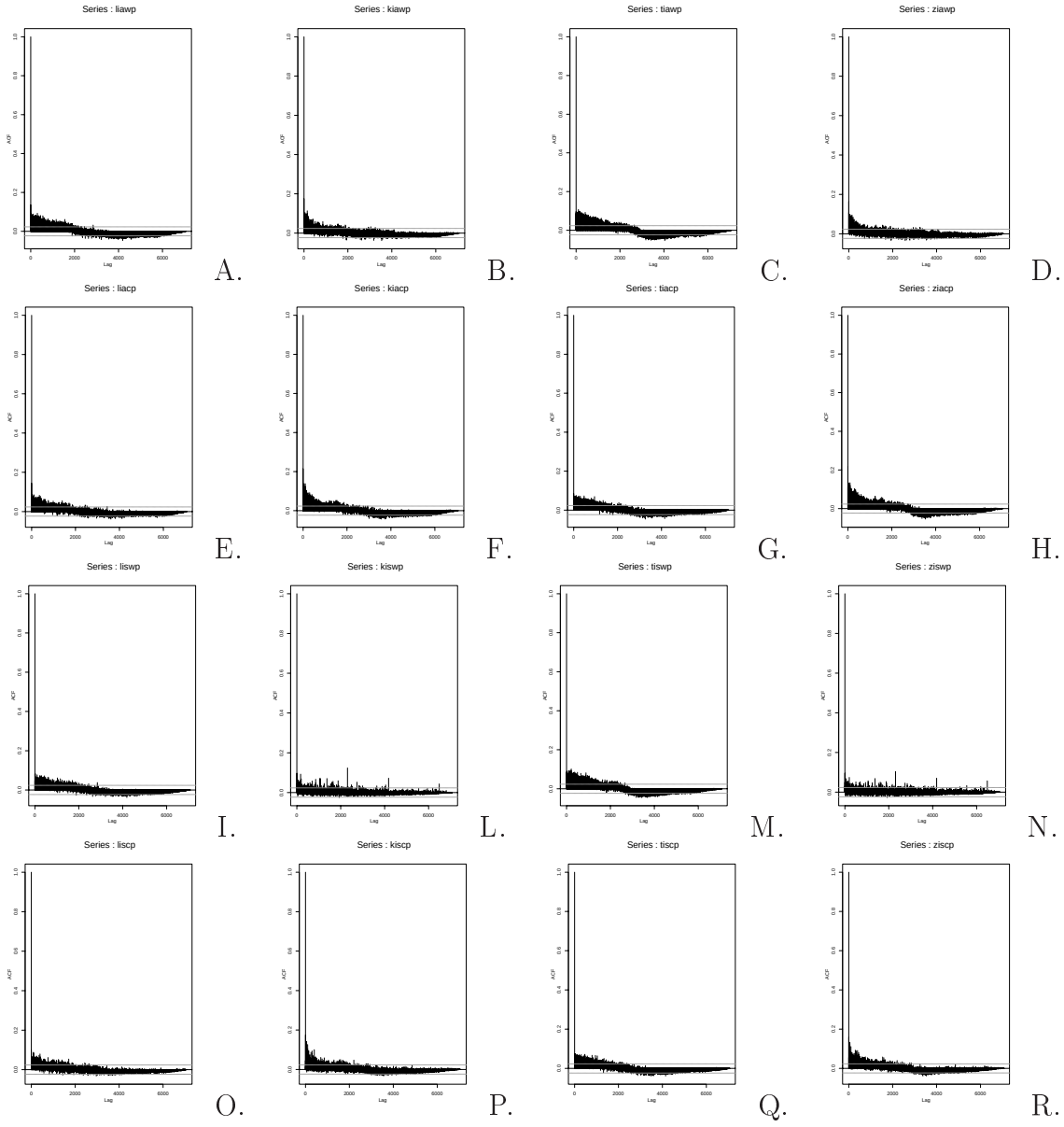


Figure 6: ACF of absolute 5m residuals from MP with 200 (A-B/E-F) and 500 (C-D/G-H) atoms on WP/CP for respectively original (first and third columns) and de-noised (second and fourth columns) tables. The same correspondent plots for squared values at (I-L/O-P) and (M-N/Q-R).

sensor signals are those computed at each resolution levels by the WP and the CP transforms. These already de-seasonalized 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 compared to 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.

Resol. lev.	0	1	2	3	4	5	6
WP-A							
level 0	**0.2218**	0.0028	0.0085	0.0047	0.0023	0.0069	0.0085
level 1	0.0002	**0.1951**	−0.0013	0.0001	−0.0189	−0.0035	−0.0037
level 2	0.0068	0.0003	**−0.167**	0.0015	0.0007	0.0019	−0.001
level 3	0.0031	−0.0057	−0.0008	**−0.1438**	−0.0019	−0.0045	0.0059
level 4	0.0012	−0.0125	0.0017	0.0028	**−0.1318**	0.0117	0.0
level 5	0.0032	−0.0023	0.0014	−0.0045	0.0008	−0.0011	**−0.1147**
level 6	0.0023	−0.0009	−0.0018	0.0047	−0.0082	**−0.121**	0.0017
CP-A							
level 0	0.0029	0.0062	0.008	0.0031	0.0021	**0.1261**	0.0033
level 1	0.0012	0.0033	0.0013	0.0013	0.0041	0.0039	**−0.1204**
level 2	−0.0089	−0.0023	−0.0031	**−0.1712**	0.0031	0.0057	−0.0006
level 3	**0.1868**	−0.0008	−0.0038	−0.0114	−0.0057	−0.0031	0.006
level 4	−0.0022	**0.1832**	0.0011	−0.0002	−0.0191	−0.0083	0.0053
level 5	0.006	0.0142	−0.0059	0.002	**0.1482**	−0.0053	0.0035
level 6	0.0014	0.0046	**0.1748**	−0.0021	0.0036	−0.0052	0.002

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

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. Thus, in this case there isn't a precise selection order, but instead low and high frequency information content is collected at various resolution degrees.
In Figure 7 we repeat the diagnostic ACF plots already shown before, based on the new residuals, and computed for their absolute and squared ACF from the the WP and CP tables.

We observe that with the WP table the dependencies left in the ACF plots are less evident than before, particularly with regard to the long memory component, while the initial autocorrelation decreases with T (the number of approximating structures). For the CP Table there is even a better ability of MP to capture and remove these dependencies, thus suggesting that in both cases the feature detection power improves qualitatively and with computational savings by simply concentrating the MP activity only on the finest resolution levels. In the WP case is the information content of the high-scale signals that

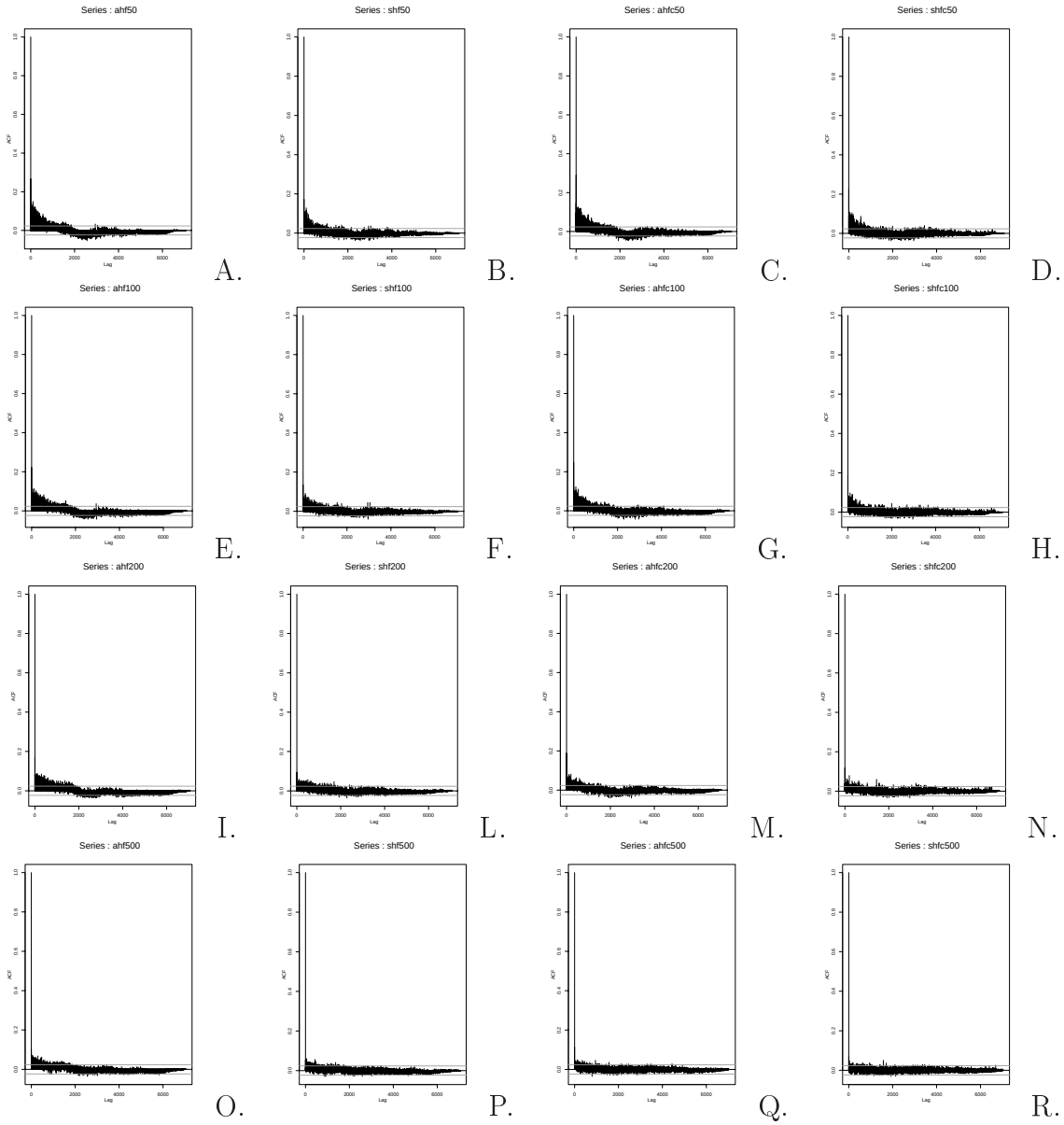


Figure 7: ACF of absolute (A,E,I,O) and squared (B,F,L,P) 5m residuals from MP with 50/100/200/500 atoms on the WP Table, with MP running on the finest four resolution levels. The correspondent plots for the CP Table are (C,G,M,Q) and (D,H,N,R), respectively.

reproduces the frequency content belonging to the low-scale ones, while in the CP case is the diagonalization power of the building blocks, i.e. the localized cosines, to make a difference.
The advantages of working with band-pass filtered detail signals in terms of temporal aggregation effects are known (Johnstone & Silverman, 1997; Abry et al., 1998); they are stationarized and de-correlated by wavelets, as seen, in the sense of being almost uncorrelated along individual scales and almost independent across scales. We support these effects with the selection operated by ICA on the wavelet expansion coefficients, justified on the grounds that the least dependent components lead to a more orthogonalized MP and thus better efficiency.

6. Locally Approximating Stock Index Volatility

One could achieve an indirect estimate of the mixed volatility by computing an average of the squared residual values obtained by the MP approximation; for pointwise volatility estimates though, another possibility is to estimate a GARCH or SV model directly on these residuals (Capobianco, 1999b).
Following (Donoho et al., 1998), with a non-stationary process an oracle orthonormal basis B would diagonalize the covariance matrix , , and thus in practice one should estimate the covariance function by rotating into the basis B, forming the empirical covariance function and then eliminating the off-diagonal terms (according to the expected values based on theoretical operators), thus getting $\Delta_B = diag(\sigma^2_{i,B})$.
Then, after rotating back in the original basis, the resulting covariance estimate is $C_B = B\Delta_B B'$. The table of empirical variances σ_i^2 is built from CP coefficients estimates, thus in the sequence space; they are smoothed by thresholding and the inverse CP transform is taken, so to get a smoothed estimate $\bar{\sigma}_i^2$. The BOB applied to a cost function built from the squared values of the smoothed empirical variances yields the basis to be used for forming the diagonal matrix $\Delta_{\bar{B}}$ and the covariance $C_{\bar{B}}$ estimates.
Our construction is instead conducted through ICA, with whitening (decorrelation) and rotation (change of basis) steps equivalently performed, respectively, in the expansion packet coefficients domain and in the MRA detail signal domain. The main difference is that instead of working directly with the covariance operator in its standard representation for stationary stochastic processes, the estimation system here introduced works more indirectly, through the MP residuals. Then, compared to the former strategy, in our procedure the focus is also on detecting the time varying variance features, and thus on the stochastic volatility or conditional heteroscedastic structure underlying the return series. The main difference in the results is that with the MP algorithm we are able to approximate the latent volatility structure with an improved localization power compared to the BOB procedure.
It is useful at this point to verify how the residues obtained by the approximation scheme contribute to the structure of volatility. We thus show parametric estimates from a model designed for the series of 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:

$$y_t = k + \xi_t \sigma_t, \quad with \;\; \sigma_t = \sqrt{h_t} \tag{6.1}$$

$$h_t = a_1 h_{t-1} + b_1 y_{t-1}^2 \tag{6.2}$$

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:

$$l_i(\Theta) = logL_T(\Theta) = \sum_{i=1}^{T} logp(y_i \mid \Psi_{i-1}) = -\frac{1}{2}\sum_{i=1}^{T} logh_i + \sum_{i=1}^{T} log \;\; g(\frac{x_i}{h_i^{1/2}}) \tag{6.3}$$

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 $x = y_i h_i^{-\frac{1}{2}}$, i.e. the standardized residuals, and that the degrees of freedom are estimated with the other parameters, say Θ, in the model).
We then check diagnostic and statistical properties, reported in Table 2 (*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 further 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, together with the ACF of the squared residuals (see Figure 8).

With 500 iterations the distributional properties suggested by the QQ-plots improve, for a better alignment with the ideal benchmark distributions. Then Figure 9 and Figure 10 illustrate a comparison between squared residual series, and the volatility estimated by the GARCH model after 100 and 500 iterations of the algorithm run on the two overcomplete dictionaries.

Here there are some interesting aspects. Result indicate that with 500 iterations the estimated volatility curve is smoother than with 100 iterations; the possibilities are that it is either approximating the true volatility or that the MP algorithm is overfitting. For

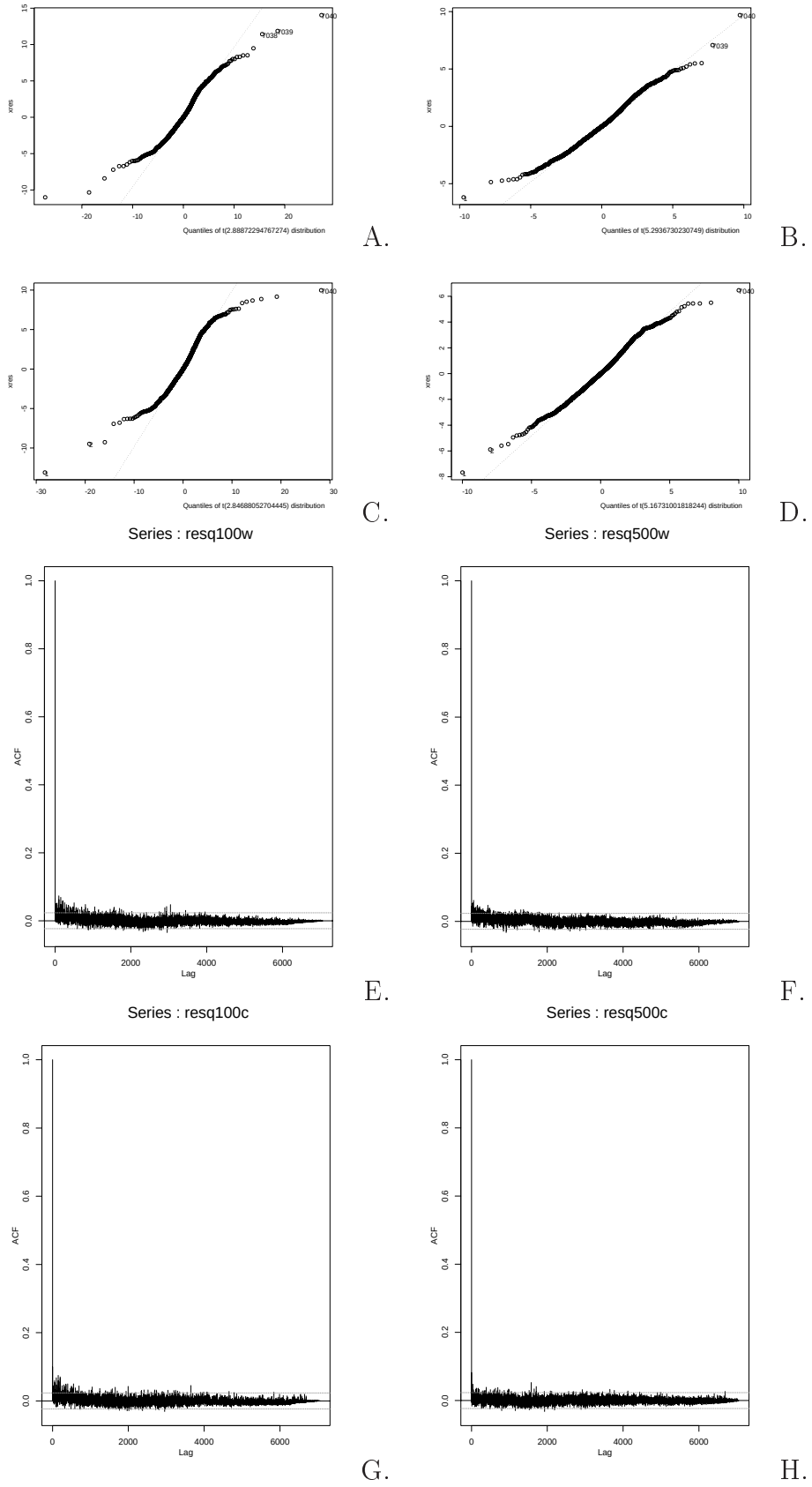


Figure 8: QQ-plots of standardized residuals vs quantiles of a Student's t, for 100 (A,B) and 500 (C,D) MP iterations run on respectively WP/CP Tables; correspondent ACF of squared residuals for WP (E,F) and CP (G,H).

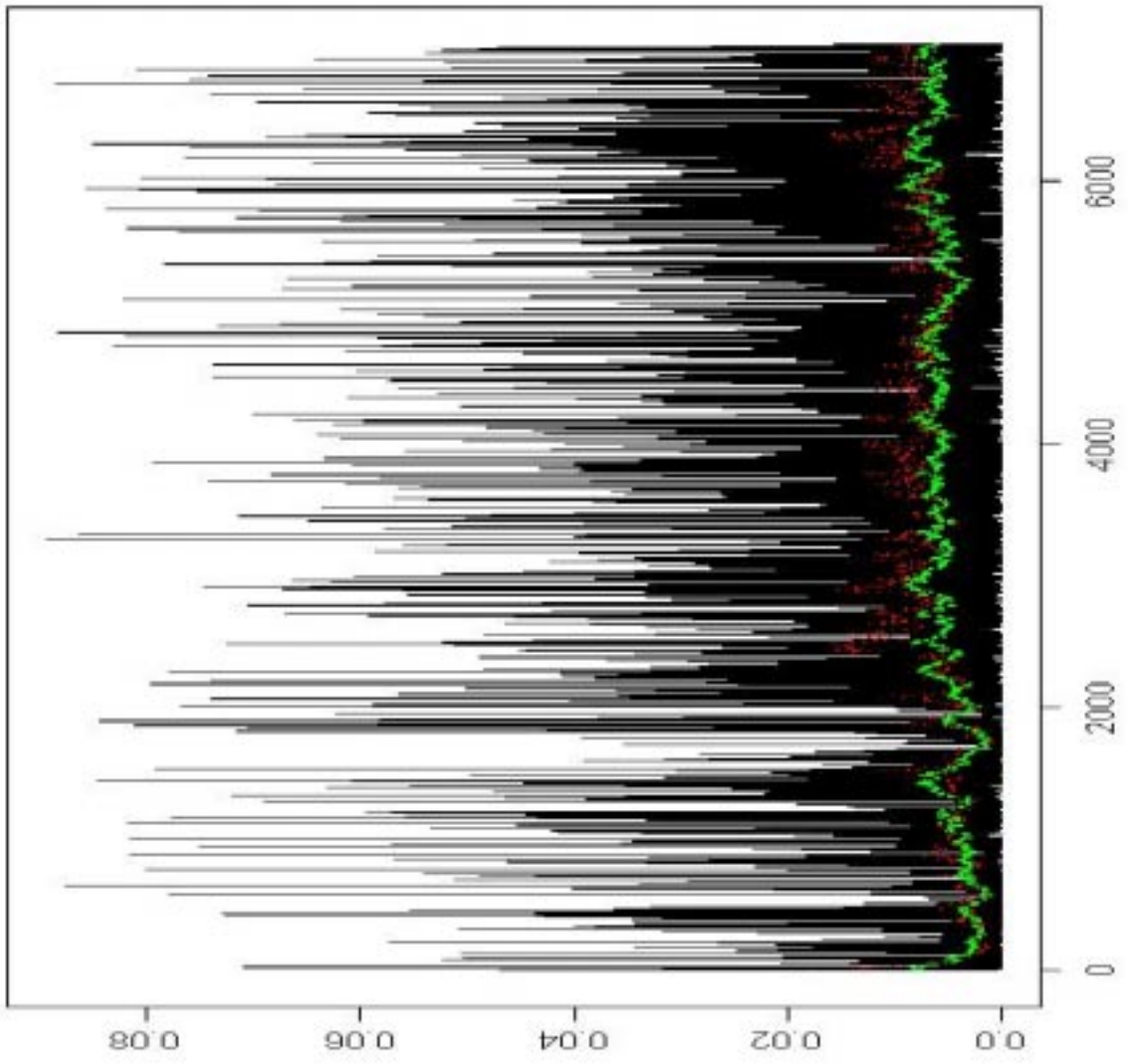


Figure 9: Squared residuals after 500 MP runs and estimated GARCH volatilities with residual series after 100 (red) and 500 (green) algorithm iterations, with WP (A).

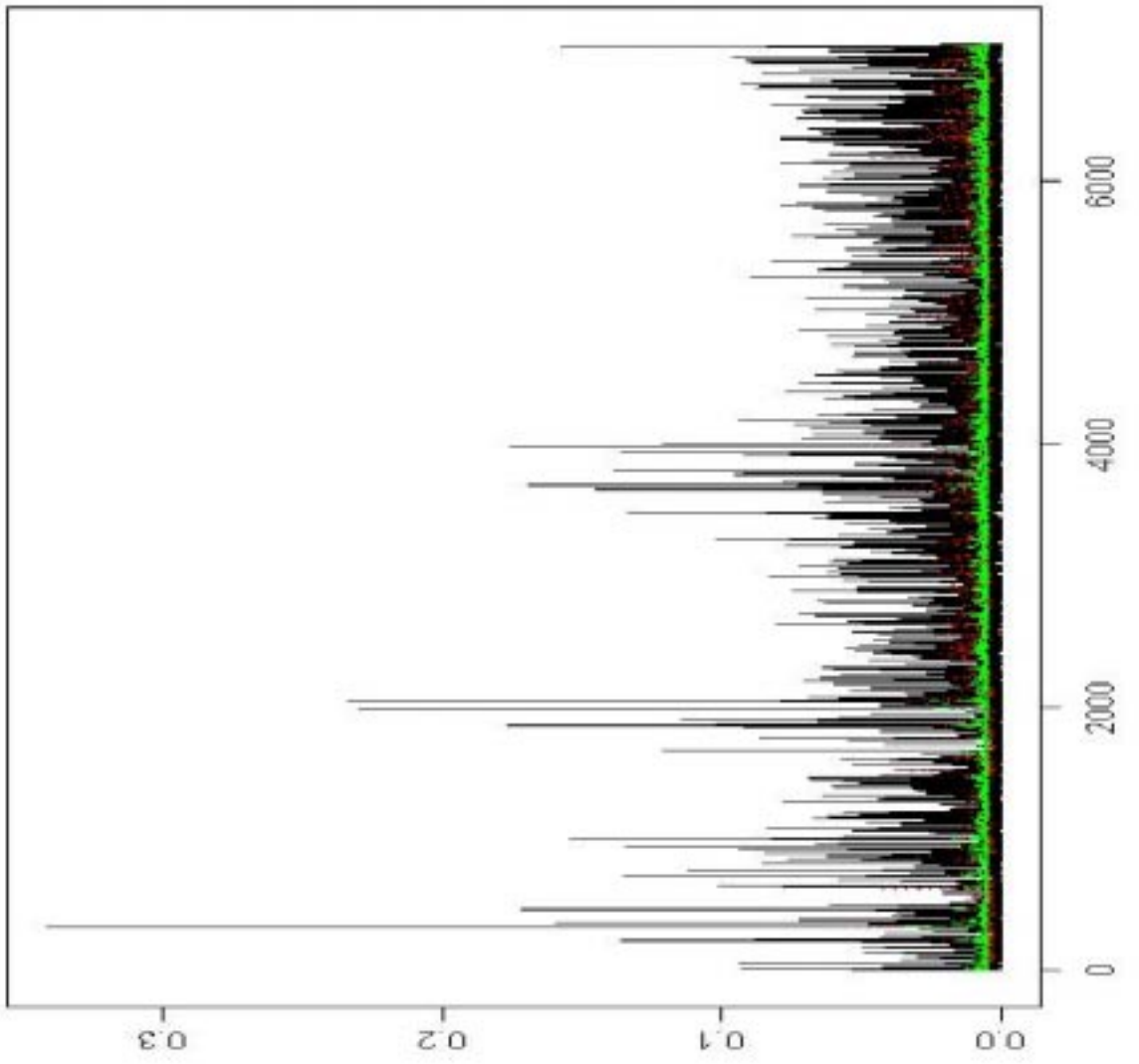
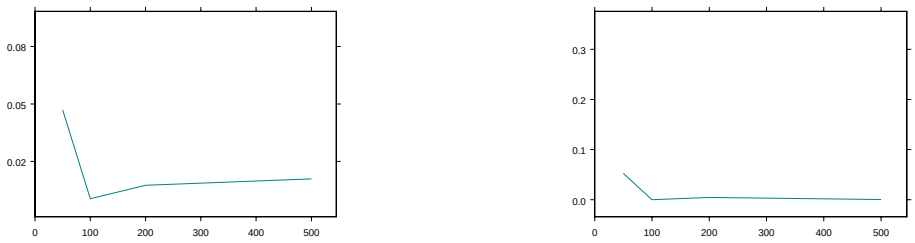


Figure 10: Squared residuals after 500 MP runs and estimated GARCH volatilities with residual series after 100 (red) and 500 (green) algorithm iterations, with CP (A).



A. B.

Figure 11: L_2 errors vs the number of approximating structures, for WP (A) and CP (B), respectively.

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. As a further test, we show in Figure 11 a comparison of L_2 and L_0 norms, with the L_0 norm representing the number of approximating structures employed by the MP algorithm.

With both the tables the MP has the best performance after 100 iterations, with a turning point in the direction of the curves at approximately 200 structures. While for the CP case the L_2 curve is smooth, for the WP case the minimum is still reached at approximately 100 atoms, but now this limit value is not followed by minimum reverting behaviour as with CP. For both dictionaries the algorithm thus reaches its best performance after 100 iterations, and it changes its pattern when more structures are included. Then, it stabilizes for CP after 200 atoms, but not for WP.
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 of volatility and at the same time for preventing MP from possible numerical instability.
This indirectly confirms that with 500 iterations, MP oversmooths the volatility function in the WP case and de-volatilizes, in fact, the return process in the CP case.

7. Conclusions

For financial time series it may be useful to adopt sparse and independent representations and decompositions. An example is offered by ICA applied to wavelet and cosine packets; in terms of feature detection power of the latent volatility structure by greedy approximation schemes such as the MP, the results we have obtained are due to the combination of goals such as achieving a sparse signal representation together with reducing the signal dimensionality through a set of least dependent coordinates over which to base the approximation.
Combining ICA with wavelet-based signal decomposition yields a sort of SCA. For non-Gaussian data one finds that wavelet-represented signals result the least dependent components selected, where the detail sequences are obtained by sequential application of WPT/CPT and ICA. These sequences are sparse, for the choice of ad hoc dictionary selection and for the packet coefficient thresholding stage.
The selection of high scale signals eliminates redundant information by keeping highly localized time resolution power without simultaneously losing too much frequency resolution, due to the fact that low scale information can be reproduced by averaging high

scales, particularly with wavelet packets.
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 MP algorithm may be very effective by just limiting its range of activity, in this case the domain of resolution levels obtained from previous signal decomposition. Exploiting the independent information content of MRA signals, as indicated by the ICA stage, may represent an efficient procedure and a near-optimal way of tuning the resolution pursuit.

Acknowledgements.

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

References

1. Abry, P., Flandrin, P., Taqqu, M.S., & Veitch, D., Wavelets for the analysis, estimation and synthesis of scaling data. In: Park, C. & Willinger, W. (eds), *Self-similar network traffic and performance evaluation*. New York: Wiley 2000, pp 39-88.
2. Abry, P., Veitch, D., & Flandrin, P., Long range dependence: revisiting aggregation with wavelets. *J. Time Ser. An.*, 19(3), 253-266 (1998).
3. Andersen, T., & Bollerslev, T., Intraday periodicity and volatility persistence in financial markets. *J. Emp. Fin.* , 4, 115-158 (1997a).
4. Andersen, T., & Bollerslev, T., Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns. *The J. Fin.* LII(3) 975-1005 (1997b).
5. Benassi, A., Locally Self Similar Gaussian Processes. In: Antoniadis, A., & Oppenheim, G., (eds), *Wavelets and Statistics*. New York: Springer-Verlag 1995, pp. 43-54.
6. Bollerslev, T., Generalized Autoregressive Conditional Heteroskedasticity. *J. Econometrics*, 31, 307-327 (1986).
7. Breidt, F.J., Crato, N., & de Lima, P., On the detection and estimation of long memory in stochastic volatility. *J. of Econometrics*, 83 325-348 (1998).
8. Bruce, A. & Gao, H.V., *S+Wavelets*, Seattle: StaSci Division, MathSoft Inc. 1994.
9. Buhlmann, P., & McNeil, A., Nonparametric GARCH models. Preprint, ETH Zurich (1999).
10. Capobianco, E., Statistical Analysis of Financial Volatility by Wavelet Shrinkage. *Meth. Comp. Appl. Prob.*, I(4) 423-443 (1999a).
11. Capobianco, E., Wavelets for High Frequency Financial Time Series. In *Interface '99 Proc.* 373-378 (1999b).
12. Cardoso, J., Source separation using higher order moments. In *Proc. Int. Conf. Ac., Sp. and Sig. Proc.* 2109-2112 (1989).
13. Cardoso, J., & Souloumiac, A., Blind beamforming for non-Gaussian signals.

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

14. Chen, S., Donoho D., & Saunders, M.A., Atomic Decomposition by Basis Pursuit. *SIAM Rev.*, 43(1) 129-159 (2001).
15. Cheng, B., & Tong, H., K-stationarity and wavelets. *J. of Stat. Plan. and Inf.*, 68 129-144 (1998).
16. Coifman, R., & Wickerhauser, V., Entropy-based algorithms for best basis selection. *IEEE Tr. Inf. Th.*, 38 713-718 (1992).
17. Comon, P.,. Independent Component Analysis - a new concept?. *Sig. Proc.*, 36(3) 287-314 (1994).
18. Dahlhaus, R., Fitting time series models to nonstationary processes. *The Ann. of Stat.*, 25 1-37 (1997).
19. Daubechies, I., *Ten Lectures on wavelets*. Philadelphia: SIAM 1992.
20. Davis, G., Mallat, S., & Avellaneda, M., Greedy adaptive approximations. *J. of Constr. Approx.*, 13(1) 57-98 (1997).
21. Donoho, D., Unconditional Bases are Optimal Bases for Data Compression and for Statistical Estimation. *Appl. Comp. Harm. Anal.*, 1 100-115 (1993).
22. Donoho, D., Unconditional Bases and Bit-Level Compression. *Appl. Comp. Harm. Anal.*, 3 388-392 (1996).
23. Donoho, D., Sparse Components of Images and Optimal Atomic Decompositions. *Constr. Approx.*, 17 353-382 (2001).
24. Donoho, D., & Johnstone, I.M., Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika*, 81 425-455 (1994).
25. Donoho, D., & Johnstone, I.M., Adapting to unknown smoothness via wavelet shrinkage. *JASA*, 90 1200-1224 (1995).
26. Donoho, D., & Johnstone, I.M., Minimax Estimation via Wavelet Shrinkage. *The Ann. of Stat.*, 26 879-921 (1998).
27. Donoho, D., Mallat, S., & von Sachs, R., Estimating Covariances of Locally Stationary Processes: Consistency of Best Basis Methods. In: *Proc. IEEE Time Freq. and Time-Scale Symp.*, 337-340. New York: IEEE 1996.
28. Donoho, D., Mallat, S., & von Sachs, R., Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods. Tech. Rep. 1998-517, Stanford University (US) 1998.
29. Engle, R.F.,. Autoregressive Conditional Heteroscedastic models with estimates of the variance of th UK inflation. *Econometrica*, 50 987-1007 (1982).
30. Fouque, J.P., Papanicolaou, G., & Sircar, R.K., Mean Reverting Stochastic Volatility. *Int. J. of Theor. and Appl. Fin.*, 3(1) 101-142 (2000).
31. Friedman, J.H., & Stuetzle, W., Projection Pursuit Regression. *JASA*, 76(376) 817-823 (1981).
32. Gao, H.Y., Wavelet shrinkage estimates for heteroscedastic regression models. Tech. Report, MathSoft Inc. 1997.
33. Ghysels, E., Harvey, A.C., & Renault, E., Stochastic Volatility. In: *Handbook of Statistics*, 14. Amsterdam: North Holland, 1996 pp.119-191.

34. Giraitis, L., Kokoszka, P., & Leipus, R., Stationary ARCH models: dependence structure and Central limit Theorem. *Econometric Theory*, 16 3-22 (2000).

35. Hafner, C.M., Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models. *J. of Stat. Pl. and Inf.*, 68 247-269 (1998).

36. Hardle, W., Kerkyacharian, G., Picard, D., & Tsybakov, A., *Wavelets, Approximation, and Statistical Applications*, New York: Springer-Verlag 1998.

37. Hastie, T.J., & Tibshirani, R.J., *Generalized Additive Models*. London: Chapman & Hall 1990.

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

39. Johnstone, I.M., & Silverman, B.W., Wavelet threshold estimators for data with correlated noise. *J. Roy. Stat. Soc., B.*, 59 319-351 (1997).

40. Jutten. C., & Herault, J., Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture. *Sig. Proc.*, 24 1-10 (1991).

41. Krim, H., & Pesquet, J.C., On the statistics of Best Bases criteria. In: Antoniadis, A., & Oppenheim, G., (eds), *Wavelets and Statistics*. New York: Springer-Verlag 1995, pp. 193-207.

42. Lewicki, M.S., & Sejnowski, T.J., Learning Overcomplete Representations. *Neur. Comp.*, 12 (2) 337-365 (2000).

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

44. Mallat, S., Papanicolaou, G., & Zhang, Z., Adaptive Covariance Estimation of Locally Stationary Processes. *The Ann. of Stat.*, 26(1) 1-47 (1998)

45. Mandelbrot, B.B., Calvet, L., & Fischer, A., The Multifractal model of asset returns. *Disc. Pap. 1164*, Cowles Foundation, Yale University, CT (1997).

46. Meyer, I., *Wavelets: algorithms and applications*. Philadelphia: SIAM 1993.

47. Mikosch, T., & Starica, C., Limit theory for the sample autpcorrelations and extremes of a GARCH(1,1) process. *The Ann. of Stat.*, 28(5) 1427-1451 (2000).

48. Neumann, M.H., & von Sachs, R., Wavelet Thresholding: beyond the Gaussian I.I.D. situation. In: Antoniadis, A., & Oppenheim, G., (eds), *Wavelets and Statistics*. New York: Springer-Verlag 1995, pp. 301-329.

49. Ray, B.K., & Tsay, R.S., Long Range Dependence in Daily Stock Volatilities. *J. Bus. Ec. Stat.*, 18(2), 254-262 (2000).

50. Serroukh, A., Walden, A.T., & Percival, D,B., Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *JASA*, 95(449) 184-196 (2000).

51. Taylor, S.J., *Modelling Financial Time Series*. Chichester: Wiley 1986.

52. von Sachs, R., & MacGibbon, B., Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scand. J. of Stat.*, 27 475-499 (2000).

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