Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

On multi-scaling in volatility processes

E. Capobianco

REPORT PNA-R0217 JUNE 30, 2002

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

On Multi-scaling in Volatility Processes

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

ABSTRACT

Among the statistical features of financial volatility processes, the most challenging for statistical inference purposes are non-gaussianity and non-stationarity. In this study I present results from experiments aimed to empirically modeling return generating processes in which the underlying volatility dynamics are subject to changes in scaling regimes. This means that the usually observed power law behavior of self-similar and long range dependent processes might be subject to dynamics which vary with the resolution levels. The observed information are encompassed in a model based on scale-dependent information packets that reflect heterogeneous activity in the financial market. I show how the volatility structure of a stock return index can be cast in a latent variable form and suggest how this model can account for regime switching when risk and returns are explicitly linked. To this end, I find that wavelets and multi-resolution analysis are very useful in modeling volatility and in interpreting its structure, since they deliver both sparse representations and coherent decompositions, and show the multi-scaling features inherent to the data set.

2000 Mathematics Subject Classification: 60H30, 62M10, 62G07.
Keywords and Phrases: Hammerstein Series; Multi-resolution; Wavelets; Reproducing Kernels; Atomic Dictionaries: Greedy Approximation; Volatility Scaling Regimes.

Note: This work was supported by ERCIM.

1. Introduction

Wavelet-based approximation and non-linear estimation techniques have been a widely studied research topic in many scientific areas, including statistics and economics. It is key, for applied work related to time series analysis, that wavelets have been found to deal well with correlated and long-range dependent series, under stationary and non-stationary conditions for the stochastic processes under investigation.

Some classes of latent variable processes, for instance those driven by volatility, show evidence of covariance non-stationarity. When this condition is not detected by standard models, statistical inference may result quite misleading. Wavelet transforms can be seen as projection operators that yield decorrelated and stationarized sequences of computed coefficient, to an extent which depends on the degree of heteroskedasticity and non-Gaussianity characterizing the observed realizations.

Financial volatility models, with both the *Generalized Autoregressive Conditional Het-*

eroskedasticity (GARCH) (Engle, 1982) and the *Stochastic Volatility* (SV) (Taylor, 1994; Ghysels *et al.*, 1996) model classes, have been popular for two decades, producing an exceptional impulse to financial econometrics.

It is crucial in finance to understand what relations exist between market returns and risk, so that individuals and institutions may optimize their investment decision process. This requires building up a model where the conditional mean and variance are separately formulated but simultaneously co-acting on the systematic component of the observation model.

Stochastic processes that allow for nonlinear dynamics and non-stationarity induced by endogenous causes or changes in regimes have attracted great attention. Interestingly, possible effects of misspecifying the model selected for the analysis should be considered, being this a limitation endowed with every parametric model.

The strategy of this paper is to shift from a standpoint of classical parametric models to a non-parametric approach for modeling complex system dynamics. Thus, we move from statistical settings relying on assumptions about the probability functions and distributions involved, to models based on greedy function approximation and semi-parametric estimation techniques.

One goal in this paper is to represent stock index return dynamics sparsely. This can be done by setting global and local function optimizers to run over atomic dictionaries and by combining multiscale methods with decomposition techniques so to achieve methodologically useful and interpretable representations for non-stationary processes.

This paper is organized as follows. In Section 2 volatility models are reviewed and then cast, in Section 3, in a linear latent variable problem setting. In Section 4, Hammerstein series and multiscale approximation systems based on wavelet techniques are described. In Section 5 decomposition techniques are introduced, while atomic dictionaries and overcomplete systems are illustrated in Section 6, together with techniques for greedy approximation. Experiments are reported in Section 7 and then, Section 8 concludes the paper.

2. The Volatility Setting

2.1 Linear Models

With the aim of briefly reviewing volatility models, we present a sequence of standard models, with increasingly more complex parametric structure. Let's start with the simple GARCH(1,1) (Bollerslev, 1986; Taylor, 1986) model with exogenous effects in mean; the observed returns y_t and the related volatility process h_t are given by:

$$y(t) = bx(t) + \epsilon(t); \qquad h(t) = \alpha_0 + \alpha_1 \epsilon^2(t-1) + \beta_1 h(t-1) \tag{2.1}$$

given $\epsilon(t)/\mathcal{F}_{t-1} \sim N(0, h(t))$[1]. The α_i coefficients must be non-negative in order to satisfy the non-negativity constraint required by the conditional variance $h(t)$[2].

In the alternative SV class of models, a stochastic mechanism is introduced, i.e., a random and independent (from the past information) shock that drives the volatility process, together with other predeterminated variables. Now h_t is no longer observable, due to the random variability coming from the separate noise. In Andersen (1994) a

[1] $\mathcal{F}_{t-1}$ is the information available through the observed data up to time $t-1$.

[2] The volatility process is not considered stochastic so far, but it changes according to the past information set $\mathcal{F}_{t-1}$.

GARCH(1,1) was generalized to a *Stochastic Autoregressive Volatility* (SARV) model in the following way.

Consider $\epsilon(t) = \sqrt{h(t)}z(t)$, with $z(t) \sim i.i.d.(0,1)$, and a volatility process given by:

$$h(t) = w + \beta h(t-1) + \alpha h(t-1)u(t) \tag{2.2}$$

with $u(t) \sim i.i.d.(1, \sigma_u^2)$, $u(t) > 0$, $z(t)$ and $u(t)$ independent from each other and with respect to $\mathcal{F}_{t-1}$.

When the volatility equation is written as $h(t) = w + \beta h(t-1) + [\gamma + \alpha h(t-1)]u(t)$, replacing $z^2(t-1)$ for $u(t)$ and given $\gamma = 0$, a GARCH(1,1) is easily obtained. Now $h(t)$ is measurable with respect to $\mathcal{F}_{t-1}$ and the model results conditionally heteroskedastic. Clearly enough, randomness of different nature in the volatility process itself complicates statistical inference, particularly when non-Gaussianity is involved.

It has been widely proposed the use of the Kalman filter algorithm in order to estimate a model cast in a state space representation; it's well known that a likelihood function can be obtained via the prediction error decomposition and the same function must be maximized with respect to the parameters of interest. When non-Gaussian errors are considered, non-linear filtering and simulation techniques must be adopted, thus needing more computational efforts (Durbin and Koopman, 2000).

2.2 Non-linear Models

ARCH-type models have also clear connections with ARMA and bilinear stochastic processes (Weiss, 1986; Guegan, 1990); in the latter case, it appears difficult to try to separate the dynamics involved in the first two conditional moments of the distribution of interest, and therefore misspecification can arise when ARCH and bilinearity are both present in a model such as:

$$\phi(B)(y_t - \mu) = \theta(B)\epsilon_t + \sum_{i=1}^{P}\sum_{j=1}^{Q}\beta_{ij}y_{t-i}\epsilon_{t-j} \tag{2.3}$$

$$h_t = \alpha_0 + \sum_{i=1}^{R}\alpha_i\epsilon_{t-i}^2 + \delta_0(\hat{y}_t - \mu)^2 + \sum_{j=1}^{S}\delta_j(y_{t-j} - \mu)^2 \tag{2.4}$$

where $E(\epsilon_t/\mathcal{F}_{t-1}) = 0$ and with $\hat{y}_t$ as the forecast for y_t calculated at time $t-1$.

Guegan (1990) considered a state space formulation for a bilinear process obtained from an ARCH model. Starting from $y(t) = \epsilon(t)h^{\frac{1}{2}}(t)$, where $h(t) = \alpha_0 + \alpha_1(y^2(t-1) - \alpha_0) + \ldots\ldots + \alpha_q(y^2(t-q) - \alpha_0)$ and given $y^2(t) = \epsilon^2(t)h(t)$, with $\epsilon^2(t) = \eta(t-1) + 1$ and $y^2(t) = x(t) + \alpha_0$, after some calculations it is easy to obtain the bilinear state space representation:

$$y^2(t) = x(t) + \alpha_0; \qquad x(t) = \sum_{i=1}^{p}\alpha_i x(t-i) + \sum_{i=1}^{p}\alpha_i x(t-i)\eta(t-1) + \alpha_0\eta(t-1) \tag{2.5}$$

where $\eta(t)$ is distributed as a χ_1^2 random variable which results not centered and having a variance equal to 2.

With the state space framework and the Kalman filter implemented on it, the estimate of the two equations can be executed in two sequential steps at each observation in the available sample. First, at time t, an estimate of the conditional variance $h(t)$ is obtained and a new innovation value $\eta(t)$ can be computed from the Kalman filter; then, this last value is used to calculate the updated quantity $h(t+1)$ and the process is repeated as before, at each observation.

More general models introduce non-linear features and usually non-parametric statistical inference procedures are needed. Let us consider Volterra series, for instance, with a representation like:

$$\begin{aligned} Y(t) &= \sum_{\tau_1=-\infty}^{\infty} h_1(\tau_1)X(t-\tau_1) + \\ &\sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} h_2(\tau_1,\tau_2)X(t-\tau_1)X(t-\tau_2) + \ldots \end{aligned} \tag{2.6}$$

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

An Hammerstein series representation improves things (Ralston *et al.*, 1997), and has the following form:

$$Y(t) = \sum_{\tau=-\infty}^{\infty} g_1(\tau)X(t-\tau) + \sum_{\tau=-\infty}^{\infty} g_2(\tau)X(t-\tau)^2 + \sum_{\tau=-\infty}^{\infty} g_3(\tau)X(t-\tau)^3 + \ldots \tag{2.7}$$

with the functions $g(.)$ as time invariant Hammerstein kernels, offering advantages in modeling separately the memory terms, and thus dealing with the non-linear system's dynamics.

The advantages are a reduction in computational load, a better mathematical tractability and a simplified system identification problem. We will adaptat this series representation so to allow for an approximation and estimation algorithm to work effectively.

2.3 GARCH-m

Our candidate model for the planned experiments is the *Generalized Autoregressive Conditional Heteroskedastic with effects in Mean* (GARCH-m). We stress the importance of the relationship between market risk and expected returns, crucial in finance theory, and take the risk premium into account in this model by allowing for the introduction of the volatility process in the conditional mean equation (Engle *et al.*, 1987).

The GARCH(1,1)-m model, following (Chou *et al.*, 1992), can be represented as:

$$y(t) = b(t)h(t) + e(t); \qquad b(t) = b(t-1) + v(t); \qquad h(t) = a_0 + a_1\eta^2(t-1) + a_2 h(t-1) \tag{2.8}$$

where $\eta(t) = y(t) - E_{t-1}[y(t)]$, $e(t)$ and $v(t)$ represent zero-mean uncorrelated white noises whose variances are respectively $h(t)$ and $Q(t)$, $h(t)$ measures the volatility and $b(t)$ is the so-called volatility price[3], while the market risk premium is $b(t)h(t)$.

We start our analysis by presenting a generalized specification of GARCH-m processes through unobserved component models. Then, we address the possible existence of volatility regimes, i.e., regimes characterized by different scaling structure of volatility.

There are results in this study which confirm the relevance of applications proposed recently by Capobianco (2002a,b) and in part based on seminal work by Abry *et al.* (1998, 2000) and Arneodo *et al.* (1998), where the main idea is that volatility presents itself with dynamic features changing with the resolution level or otherwise the scale at which we observe returns. Then, a multiresolution decomposition of signals like that allowed by wavelets and related families becomes very informative and useful for statistical inference purposes.

3. Latent Variable Systems

3.1 Model Specification

We set a linear representation for the observed system dynamics:

$$Y(t) = A(t)S_h(t) + \xi(t) \tag{3.1}$$

$$S_h(t) = C(t)\Phi(t) + \eta(t) \tag{3.2}$$

where $Y(t)$ are to be considered observed financial returns[4], $S_h(t)$ are unknown risk factors which independently drive the volatility process based on an atomic function dictionary representation and dynamics $\eta(t)$, $A(t)$ is an unknown mixing matrix and $\xi(t) \sim i.i.d.[0, \sigma_\xi(t)]$ is a noise process.

Then, $h(t) = \sigma_\xi^2(t) = f[S_h(t)] \approx f[C\Phi]$, for f unknown, can be considered a ridge-like approximation form of the volatility process underlying the returns dynamics, which can then be reduced to a linear combination of risk factors, i.e., $h(t) \approx \sum_i w_i \mid S_{h,i}(t) \mid + \epsilon(t)$, with $\epsilon(t)$ as approximation error from the underlying non-linear relation linking volatility and risk factors. The index i refers to a certain number of risk factors influencing the volatility process and modulated by w_i.

The system encompasses GARCH specifications, after some adjustments. Matching the volatility specification, and depending on the building block structure of the Ψ_t dictionary, we can have GARCH-type models (for $\eta(t) = 0$) or SV models (for $\eta(t) \neq 0$). In both these cases we have risk factors and volatility, the latter being a global measure of risk compared to the more local one represented by individual factors. Models (2.5) and (2.8) can also be recovered as special cases.

Note that we implicitly assume a linear relation in the shape of the market risk premium as an approximation to a non-linear one. The fact that we express the underlying system

[3]The time-varying coefficient relating the first two conditional moments is assumed to be distributed as a random walk, instead of being a constant term. Thus, an additional source of randomness is introduced.

[4]Stock returns are computed in the usual way, as $r(t) = ln[p(t)/p(t-1)] \times 100$, where $p(t)$ are the prices of shares, indexes, commodities or other financial activities.

dynamics through an arbitrary functional form addresses the possibility of non-linearity underlying the system (Backus and Gregory, 1993).

Thus, let us equivalently call $S_h(t)$ sources of volatility, specified in a completely non-parametric way and such that a latent volatility process can be still specified following one of its realization[5] and conditionally on the observed return series, or otherwise according to a data generating mechanism subject to more complex stochastic dynamics. A mixture structure for the volatility dynamics remains as the main force underlying the system, under this model strategy.

The sources $S_h(t)$ have a possibly sparse decomposition through $\Phi(t)$, a selected dictionary of functions delivering a basis or an overcomplete representation (Lewicki and Sejnowski, 2000; Chen *et al.*, 2000) for the signal under investigation. The expansion coefficients are indicated by $C(t)$, while $\eta(t)$ is an i.i.d process, with no constraints on the probability distributions.

A special case (Zibulewsky and Pearlmutter, 2001) is when a dual system can be formed and a basis is obtained. In that case the system can change according to the transform $\Phi^{-1}(t) = \Psi(t)$. As a direct consequence, $S_h(t)\Psi(t) = C(t)\Phi(t)\Psi(t) + \eta(t)\Psi(t)$, which can also be expressed as $\tilde{S}_h(t) = C(t) + \tilde{\eta}(t)$. It follows that $Y(t) = A(t)C(t) + A(t)\tilde{\eta}(t) + \xi(t)$ or also $Y(t) = A(t)C(t) + \xi(t)$, with $\xi(t) = A(t)\tilde{\eta}(t) + \xi(t) \equiv A(t)\eta(t)\Psi(t) + \xi(t)$.

From these transformations, a new system is found:

$$\tilde{S}_h(t) = C(t) + \tilde{\eta}(t) \tag{3.3}$$

$$Y(t) = A(t)C(t) + \zeta(t) \tag{3.4}$$

If the *signal-to-noise ratio* (S/N) results high with regard to the stochastic nature of the sources of volatility, then $\eta(t) \approx 0$ and $\zeta(t) = \xi(t)$, implying that the same volatility process initially described is found. If instead S/N is low, the square root of the volatility process becomes characterized by $\Sigma(t) = D(t) + \sigma_\xi(t)$, where $D(t) = A(t)\sigma_\eta(t)\Psi(t)$.

As a result, and from a statistical perspective, we have the volatility in non-parametric form in (3.1-3.2), investigated by selected atomic dictionaries of functions. Alternatively, we have a system (3.3-3.4), where the mixing A_t is now acting on the computed transform expansion coefficients C_t. In other words, there is a switch from signal (function) to sequence (feature) space, i.e. with continuous and/or discrete time functional relations in the former and with sequences of expansion coefficients in the latter.

3.2 The Methodological Approach

Cascade Systems Processing We first go back to the Hammerstein series approach, and build a time-varying Hammerstein version according to:

$$Y(t) = \sum_{\tau=-\infty}^{\infty} G[X(t-\tau), t, \tau] \tag{3.5}$$

To make it effective in approximation, we might simply expand the kernels in a polynomial basis sequence that thus captures arbitrary high order system's dynamics and then correspondingly adapt the model as follows, in a signal+noise fashion:

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

$$Y(t) = \sum_{n=1}^{N} \sum_{m=-M}^{M} g_n(t,m) X(t-m)^n + \epsilon(t) \tag{3.6}$$

In order to further reduce the complexity we may adopt as Hammerstein kernels linear combinations of basis functions $\psi_k(t)$, for instance some wavelet families, harmonics as in Fourier analysis, splines, frames or other approximators. When dealing with non-stationary series it is wise to use compactly supported functions, for reasons good localization power and thus inhomogeneous smoothness adaptivity. Thus we have, accordingly, the revised model:

$$Y(t) = \sum_{n=1}^{N} \sum_{m=-M}^{M} \sum_{k=0}^{K} \beta_n(k,m) \psi_k(t) X(t-m)^n + \xi(t) \tag{3.7}$$

whose estimation can be conducted under no special conditions, rather than assuming Gaussian density fucntions and/or stationarity of random input processes. The analogous integral functional form for continuous time version (Greblicki, 2000) of the model is:

$$Y(t) = \int_{-\infty}^{\infty} \Upsilon(t-\tau) K(X(\tau)) d\tau \tag{3.8}$$

As done before, another more practical formulation is used for estimating with a time-varying, truncated, finite support form of (16):

$$Y(t) = \sum_{i=1}^{N} \sum_{\gamma=0}^{l} \Upsilon_i(t,\gamma) K_i[X(t,\gamma)] + \epsilon(t) \tag{3.9}$$

This system depends on the specification of the Υ and K dynamics, and on $f(\epsilon)$, the unknown error *pdf*. Up to this stage, the statistical model can be parametric, semi- or non-parametric, depending on our *a priori* knowledge and on the assumptions we want to use.

Given a random process $X(t)$ as the input to the cascade system, with a bounded variance and an unknown density, a non-linear subsystem $Z(t) = K(X(t))$ which is unobservable with $K(.)$ bounded and invertible, and a linear component $\Upsilon(t)$ such that $\int_0^\infty \Upsilon^2(t)dt < \infty$, the model identification conditions refer to $\int_{-\infty}^{\infty} f^2(x)dx < \infty$ together with boundedness away from zero of f so to select just high density regions and enable L^2 multiscale decomposition and estimation. This last condition is less important for estimation in inhomogeneous smoothness function spaces, where other devices will have to be used (eg., thresholding).

The conditions lead to a functional regression problem, due to the independence of the error in the model, i.e., $E[Y(t+\tau) \mid X(t) = x]$, for which standard parametric models are not applicable if we want to maintain a flexibility in the model toward properties like localization, sparsity and resolution ability. Kernel estimation and orthogonal series principles find an optimal bridge in wavelets (Hasiewicz, 2001), whose fast convergence applies for large class of input processes and adaptively for non-linearities through the thresholding devices (Donoho and Johnstone, 1994-98; DeVore, 1998).

Independent and Sparse Component Analysis Recovering the non-linear structure is the hardest part, but it becomes feasible in the Hammerstein setting via regression. However, even if possible, the recovery may be achieved up to accuracy dependent on the linear subsystem estimation. We adopt Blind Deconvolution techniques like Independent Component Analysis (ICA) (Cardoso, 1989; Comon, 1994) and Sparse Component Analysis (SCA) (Donoho, 2001). Ideally, one attempts to combine the advantages delivered by sparsity and independence of signal decomposition, which transfer to better model estimation, combined with signal compression and reconstruction power.

The goal of ICA is searching for statistically independent coordinates characterizing certain objects and signals, or otherwise for least dependent coordinates, due to a strong dependence in the nature of the stochastic processes observed through the structure of the index series. The combination of this goal with that of searching for sparse signal representations suggests hybrid forms of SCA.

By assuming that the sensor outputs are indicated by $x_i, i = 1, \ldots, n$ and represent a combination of independent, non-Gaussian and unknown sources $s_i, i = 1, \ldots, m$, a non-linear system $Y = f(X)$ could be approximated by a linear one AS, where $X = AS$. Instead of computing $f(X)$ one may now work for estimating the sources S together with the $m \times m$ mixing matrix A, where usually $m << n$, with n the number of sensor signals, but with $m = n$ holding in many cases too.

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

Independent components can be efficiently computed by ad-hoc algorithms such as joint approximate diagonalization of eigenmatrices for real signals (*JadeR*) (Cardoso and Souloumiac, 1993). For Gaussian signals, the Independent Components are exactly the known Principal Components; with non-Gaussian signals ICA delivers superior performance, due to the fact that it relies on high order statistical independence information.

Approximation Since the sources of volatility are unobservable, the goal is estimating them, together with the mixing matrix. These are quite complicated tasks. 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 recursively, in the mean square sense, through some iterations of a greedy approximation algorithm.

The idea pursued in this work is processing the observed returns with a greedy approximation and de-noising technique that works through atomic dictionaries with approximating structures which are selected to be the wavelet packet (WP) (Coifman *et al.*, 1992) libraries. Therefore, one looks at the following general approximation scheme:

$$Y_t \approx P_t \Phi_t + \xi_t = A_t C_t \Phi_t + \eta_t \tag{3.10}$$

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

Even if a greedy algorithm already works by sparsifying P_t through overcomplete representations, it cannot separate the components of the operator P_t. Another decomposition step is required and in fact we show how well ICA can deal with this aspect. While the A_t

mixing matrix accounts for modulating the dependence structure of the latent volatility sources, the WP expansion coefficients represents the inputs for the ICA step in feature space. This methodological proposal and the corresponding hybrid algorithm are novel aspects proposed in Capobianco (2002c), as far as concerns volatility studies.

Within- and Across-scale Dynamics Following Arneodo *et al.* (1998) and the analogies between price dynamics and hydrodynamic turbulence studies, multiplicative cascade models can be suggested for modeling financial return dynamics, such that $y_s(t) = \sigma_s(t)\epsilon(t)$, with $\sigma_s(t) = \sqrt{h_s(t)}$ and the scale $s < \bar{s}$, the latter being the integral scale, combined with a volatility multiplicatively decomposed over the sequence of scales s_j, for $j = 0, \ldots, J$, $s_0 = \bar{s}$ and s_J the finest resolution level, thus leading to the following cascade representation:

$$\sigma_s(t) = \Pi_{j=0}^{J-1} \lambda_{s_j} \sigma_{\bar{s}}(t) \tag{3.11}$$

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

Consider (3.2), where the system dynamics are subject to a mixing mechanism, at each scale, and the volatility is expressed as a consequence in ridge approximatiojn form. But if we look at the fixed scale s in (3.11), say λ_s, we have $\sigma_s(t) = \lambda_s \sigma_{\bar{s}}(t)$, and the complementarity between the two represented dynamics becomes more clear. Combined with a within-scale mixture structure and by varying s, an across-scale cascade structure can be thus set.

We don't pursue further, for now, this topic of random cascades, but from the model building perspective here adopted we focus on the coefficients which are now sparsely represented and investigated in separated sources of volatility information through ICA. The original returns have a new decomposition leading from the latent volatility components to transformed and scaled ones, embedded in projected detail signals. Our goal is to show how compression and decorrelation power of wavelet transforms can be supported by a decomposition in least dependent coordinates obtained with ICA.

4. Wavelets and Reproducing Kernels

4.1 The Continuous Wavelet Transform

Compared to Fourier bases and their optimality under stationary conditions, in the presence of non-stationary random processes there is a need for more local and sparsified representations. Thus, one might consider multiscale representations; wavelets (Daubechies, 1992; Meyer, 1993), for their strong localization and adaptive power, have been successfully used in non-stationary systems.

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

$$WT^c(f)_{jk} =< f, \psi_{jk} >=\mid j \mid^{-\frac{1}{2}} \int f(t)\psi(\frac{t-k}{j})dt \tag{4.1}$$

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

$$f = c_{\psi}^{-1} \int \int WT^{c}(f)_{jk} \psi_{jk} \frac{djdk}{j^2} \tag{4.2}$$

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

$$\forall f, g \in L^2(R), \quad \int \int WT^{c}(f)_{jk} \widehat{WT}^{c}(g)_{jk} \frac{dkdj}{j^2} = c_{\psi} < f, g > \tag{4.3}$$

Given a constant $c_{\psi} < \infty$ and integration range $(-\infty, \infty)$, 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 (RKHS) (Aronszajn, 1950). To see this, just replace from the resolution formula g with ψ and $K(j,k,j',k') =< \widehat{WT}^{c}(g)_{jk} > = < \psi_{j'k'}, \psi_{jk} >$.

The redundancy of the wavelet transform, implicit in the time-scale decomposition operated in the projected domain, is reflected in the existence of the RKHS, which given $f \in L^2(R)$, is the image of that space by the transform:

$$H_f = \{F \in L^2(R^2), P_f F = F\} \tag{4.4}$$

where P_f is the integral operator with reproducing kernel K_f defined by:

$$K_f(j,k,j',k') = \frac{1}{c_{\psi}} < f_{(j'k')}, f_{(jk)} > \tag{4.5}$$

or otherwise the orthogonal projection onto H_f, i.e., the operator satisfying self-adjointness and idempotence, $P_f^* = P_f^2 = P_f$.

Equivalently, an RKHS is the space of solutions $\upsilon(b,a)$ of the integral equation:

$$\upsilon(b',a') = P_{\psi}\upsilon(b',a') = \int_{-\infty}^{\infty} K_{\psi}(b',a',b,a)\upsilon(b,a)\frac{da}{a}db \tag{4.6}$$

where the reproducing kernel K_{ψ} is given by:

$$K_{\psi}(b',a',b,a) = \frac{1}{c_{\psi}} < \psi_{(ba)}, \psi_{(b'a')} > \tag{4.7}$$

Re-adapting what has been previously stated in terms of the covariance function, one finds:

$$E[WT^{c}(f)_{jk}\widehat{WT^{c}}(f)_{j'k'}] = c_{\psi} K_{\psi}(j,j',k,k') \tag{4.8}$$

Due to the correlation structure accounted by the reproducing kernel K_{ψ}, only almost eigenfunctions are provided by the wavelet bases. In this way, only an almost diagonalization can be achieved for the covariance.

4.2 Multiresolution Analysis

Consider general function estimators in the form of $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, X_i)$, with $K(.)$ positive definite and symmetric, as required by a natural RKHS setting. This estimator can be produced by the Multiresolution Analysis (MRA) (Mallat, 1989), as projection kernel in $L_2(R)$ operating through nested spaces $\ldots, V_{j-1}, V_j, V_{j+1}, \ldots, j \in Z$ that approximate the space. Thus, it holds that:

$$P_V f(x) = \int_{-\infty}^{\infty} K_\nu(x, y) f(y) dy \tag{4.9}$$

for

$$K_\nu(x, y) = \frac{1}{\nu} K(\frac{x}{\nu}, \frac{y}{\nu}) \tag{4.10}$$

and with

$$K(x, y) = \sum_{k \in Z} \phi(x - k)\phi(y - k), \quad k, j \in Z, \;\; \nu = 2^{-j} \tag{4.11}$$

where $\phi(x)$ is a wavelet scaling function and $\psi(x)$ is a mother wavelet such that their $k-$translated sequences form orthonormal bases for, respectively, the spaces V_0 and W_0, with W_j being the orthogonal complement of V_j in V_{j-1}. Thus, V_j is a RKHS with a unique reproducing kernel $K_\nu(x, y)$; the order of the vanishing moments, likewise the order of reproducibility, depend on the MRA characteristic rather than on the choice of the wavelet basis itself.

Note that a wavelet family $\psi(x)$ has vanishing moments of order M if and only if the following moment conditions hold:

$$\int_{-\infty}^{\infty} \psi(x) x^l dx = 0 \tag{4.12}$$

for $l = 0, 1, \ldots, M - 1$, with $\int_{-\infty}^{\infty} \psi(x) x^M dx \neq 0$. With regard to the key role of the vanishing moments, kernel expansions of integral operators in a wavelet basis allow the transformed entries to decay at a faster rate compared to the original kernel, where the rate depends on some measured distance from the diagonal. Possessing several vanishing moments allows for an improved diagonalization power.

The flexibility from choosing the most appropriate number of vanishing moments delivers robustness to non-stationarities, while decorrelating and stationarizing. By expanding a function over some basis, the dependence structure is learned through both the basis and the expansion coefficients; it has been shown (Abry *et al.*, 1998) that for long memory processes it is mostly up to the wavelet basis to capture dependence, leaving an almost uncorrelated sequence of expansion coefficients.

5. Approximation through Atomic Dictionaries

5.1 Non-orthogonal Systems

Inhomogeneous function classes characterization, diagonalization and denoising power, sparsity yield, together with the multiresolution property, a powerful justification for selecting wavelets as approximation and estimation techniques.

The Dyadic Wavelet Transform (WT^d) is simply a map $f \rightarrow w$ from the function (signal) domain to the representation (sequence) domain, that of wavelet coefficients. One applies a bank of quadrature mirror filters, by the system $w = Wf$, so to get the coefficients for the high and for the low scales.

This WT^d has still, as for the WT^c of which is a sampling result, a reproducing kernel operator K_ψ embedding the correlation structure and acting as a convolution of an inverse and a direct WT^d. When $F(j,k) \in L^2(Z \times R)$, the following holds (Carmona *et al.*, 1998):

$$K_\psi(j,k) = \sum_{j'} \int < \tilde{\psi}_{k'}^{2^j}, \psi_k^{2^j} > F_{j'k'} dk' \tag{5.1}$$

With $\tilde{\psi} \neq \psi$ a dual wavelet defined in the Fourier domain, such that for any $f \in L^2(R)$ and with wavelet coefficients computed as $< f, \psi_k^{2^j} >$, we see that the K_ψ action comes from the inner product of wavelet and dual wavelet transforms, the latter defining an inverse transform and thus used in the reconstruction formula:

$$f(x) = \sum_{j} \int < f, \psi_k^{2^j} > \tilde{\psi}_k^{2^j}(x) dk \tag{5.2}$$

While there are clear advantages from an orthogonal wavelet expansion, when combined with usual standard Gaussian assumptions underlying the system, and they refer to the possibility of finding independent coordinates in the decomposition domain of wavelet expansion coefficients even in the presence of correlation, other families of non-orthogonal functions have been suggested and employed.

Two main examples are the biorthogonal wavelets and the undecimated wavelet transforms (WT^u). In the former case dual sets of functions, resulting non-orthogonal within each set but orthogonal between the sets, are the building blocks of the approximation systems that find successfull examples of application in inverse problems, for instance (Johnstone, 1999).

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

$$Wd = Wf + \sigma Wz = w + \sigma\epsilon \tag{5.3}$$

where the Gaussian property is preserved, i.e., $\epsilon \sim N(0, \sigma)$, with $\sigma = WW'$, despite the presence this time of structure in $\sigma = WW' \neq I$, and of heteroscedasticity when σ is time-dependent.

5.2 Dealing with Strong Dependence

In a non-i.i.d. and non-stationary context, these aspects are even more emphasized and heteroscedasticity plays a major role, requiring a different treatment via resolution-wise

thresholding. Now $corr(\epsilon_i, \epsilon_i)$ becomes a key factor, and this correlation structure has to be accounted for by the correlation kernel, requiring a larger threshold compared to the orthogonal setting. In compact form and for the WT^u, the various correlation components of the projected correlation kernel are derived from the following equations (Berkner and Wells, 1998):

$$cov(c_k^j, c_{k+\tau}^j) = \Phi(2^{-j}\tau) \tag{5.4}$$

$$cov(d_k^j, d_{k+\tau}^j) = \Psi(2^{-j}\tau) \tag{5.5}$$

while with regard to the cross-correlation structure, i.e., the correlation across scales j, j', it holds that:

$$cov(c_k^{j'}, d_{k+\tau}^j) =< \phi_{j-j'}, \psi(. - 2^{-j}\tau) > \tag{5.6}$$

$$cov(d_k^{j'}, d_{k+\tau}^j) =< \psi_{j-j'}, \psi(. - 2^{-j}\tau) > \tag{5.7}$$

Thus the number of resolution levels for the chosen wavelet system determinates the degree of correlation transmitted by the transform in the wavelet coefficient domain. One of the most important properties of wavelets are the de-correlation and stationarizing power, visible from the expansion coefficient sequences behavior. For a stochastic process X with strong dependence structure (Abry *et al.*, 2000; Johnstone and Silverman, 1997), and given $E[d_x(j,k)] = 0$, the variance is:

$$E[d_x(j,k)^2] = \int \Gamma_x(v) 2^j \mid \Psi_0(2^j v) \mid^2 dv \tag{5.8}$$

and 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 $v \to 0$, i.e., power law behavior, then for $j \to +\infty$ and for $\alpha \in (0,1)$, $E[d_x(j,k)^2] \sim 2^{j\alpha} c_f C(\alpha, \psi_0)$, with $C(\alpha, \psi_0) = \int \mid v \mid^{-\alpha} \mid \Psi_0(v) \mid^2 dv$.

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

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 M satisfies a constraint, and the variance of $d_x(j,k)$ shows scaling behaviour in a range of cut-off values $j_1 \leq j \leq j_2$ to be determined. The sequence no longer shows long range dependence (LRD) but only short range dependence (SRD) when $M \geq \frac{\alpha}{2}$, and the higher M the shorter the correlation left, due to:

$$E[d_x(j,k) d_x(j,k') \approx \mid k - k' \mid^{\alpha - 1 - 2M}, for \mid k - k' \mid \to +\infty \tag{5.9}$$

$$E[d_x(j,k)d_x(j',k') \approx \mid 2^{-j}k - 2^{-j'}k' \mid^{\alpha-1-2M}, for \mid 2^{-j}k - 2^{-j'}k' \mid \to +\infty \qquad (5.10)$$

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 $(j,k) \neq (j,k')$ and $E[d_x(j,k)d_x(j',k') = 0$ for $(j,k) \neq (j',k')$. One can then look at the variance as follows:

$$var(d_x(j,k)) \approx 2^{j\alpha} \qquad (5.11)$$

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

5.3 Dealing with Volatility Processes

Due to the fact that daily squared returns don't help too much in forecasting the latent volatility structure, the concept of realized volatility (Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2001) the concept of realised volatility has been suggested so to obtain a more accurate estimate of the function with high frequency observations.

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

It holds, due to the quadratic variation principle (Karatzas and Shreve, 1998), a result like the following: given X and the partition $T = \{t_0, t_1, \ldots, t_n\}$ of $[0,t]$, the p^{th} variation of X over T is $V_t^{(p)}(T) = \sum_{k=1}^{n} \mid X_{t_k} - X_{t_{k-1}} \mid^p$. Then, if $\mid\mid T \mid\mid = max_{1\leq k\leq n} \mid t_k - t_{k-1} \mid$ goes to 0, for $p = 2$, the case of interest here with volatility, we have $lim_{\mid\mid T \mid\mid \to 0} V_t^{(2)}(T) =< X >_t$, where the limit is the quadratic variation of X.

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

We face two problems when approximating the realized volatility function and estimating the model parameters involved: the role of smoothness and the presence of noise. Following Antoniadis and Lavergne (1995), we might rely on quadratic information from the data, leading to non-negative estimators of the following kind:

$$\tilde{\sigma}^2(t) = \sum_i \alpha_i r_i^2(t) \qquad (5.12)$$

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:

$$\sigma_W^2(t) = \sum_k < \sigma^2(t), \phi_{j_0,k} > \phi_{j_0,k}(t) + \sum_{j>j_0} \sum_k < \sigma^2(t), \psi_{j,k} > \psi_{j,k}(t) \qquad (5.13)$$

where a smooth part is combined with a cumulated sequence of details obtained at different scales. We must then apply the same decomposition to $\tilde{\sigma}^2$, being σ_2 unobservable, and can obtain a perturbed version of the previous approximation:

$$\hat{\sigma}^2(t) = \tilde{\sigma}^2_W(t) + \epsilon(t) \tag{5.14}$$

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

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

5.4 Overcomplete Representations and Greedy Approximation

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

An example of overcomplete representations is offered by WP, an extension of the wavelet transform to a richer class of building block functions that allow for a better adaptation due to an oscillation index f related to a periodic behaviour in the series and deliver a richer combination of functions. 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 $\{2^{-\frac{j}{2}} W_f(2^{-j}t - k), k \in Z, (j, f) \mid I_{j,f} \in P\}$ (Krim and Pesquet, 1995).

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.

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

The MP algorithm (Mallat and Zhang, 1993) is a good example, and it has been successfully implemented in many studies for its simple structure and effectiveness. A signal is decomposed as a sum of atomic waveforms, taken from families such as Gabor functions, Gaussians, wavelets, wavelet and cosine packets, among others. We focus on the WP table, whose signal representation is given by:

$$WP(t) = \sum_{jfk} w_{j,f,k} W_{j,f,k}(t) + res_n(t) \tag{5.15}$$

In summary, the MP algorithm approximates a function with a sum of n elements, called atoms or atomic waveforms, which are indexed by H_{γ_i} and belong to a dictionary Γ of functions whose form should ideally adapt to the characteristics of the signal at hand.

The MP decomposition exists in orthogonal or redundant version and refers to a greedy algorithm which at successive steps decomposes the residual term left from a projection of the signal onto the elements of a selected dictionary, in the direction of that one allowing for the best fit. At each time step 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:

$$f(t) = \sum_{i=1}^{n} h_i H_{\gamma_i}(t) + res_n(t) \tag{5.16}$$

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 from step 2, until $i \leq n$.

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

6. Application

6.1 Data

We refer to the Nikkei stock return index with the series available, among others, for 1990, with observations collected at a very high frequency, i.e. every minute (1min). The total sample has 35,463 observations, with intra-daily trading prices covering the working week, and excluding holidays and weekends. We then form a temporally aggregated time series of correspondent five-minute (5min) data from the original one; these are simply given averaging the realization points sampled at the 1m time interval. The aggregated sample consists of 7092 observations[6]

In Figure 1 we report the time-frequency tiling given by WP decomposition (A) and the top-100 largest coefficient approximation with the MP algorithm (B), together with a sub-sample performance after iterating 50, 100, 200 and 500 times (C-D). The WP segmentation works through the initial sample split into segments[7], where their choice simply reflects the dyadic sample design, with a sample splitting rule which restricts the partition to sample sizes divisible by 2^J. We keep it at the simplest level of partitioning, and just consider two sub-samples, with observation ranges 1-3328 and 3329-7040, for the 5min series.

The plots suggest that MP works efficiently, due to its greedy nature, and a better ability to capture the local features, both in time and in frequency. The MP scheme exploits the correlation power inherent to the collection of waveforms available through the WP dictionary, and it does so throughout more scales and by extending the basis which represents the signal.

6.2 Experiments

In Figures 2 we report the extracted WP sources, obtained after 100 iterations of the MP algorithm, which is a safe amount of computing before the numerical approximator starts to overfit (Capobianco, 2002c). The sources appear for every resolution level, according to the building block characteristics of the family of symmlet-based and localized cosine packets used.

In Figure 3 the plots show the variability of the WP and source sequences across scales. With the WP sources there is an higher variability, and also more heterogeneity compared to CP. Figure 4 reports the variance of the WP sequences and sources[8] computed at each resolution level, together with the log-energy plots measuring the average energy at each scale j according to:

$$E_j = \frac{1}{N_j} \sum_{k=0}^{N_j-1} \mid WP_{j,k} \mid^2 \tag{6.1}$$

[6]The experiments have been conducted with *S+PLUS* and *Matlab*.

[7]This is one of the strategies usually adopted to tackle non-stationarity in long realizations.

[8]They are normalized, thus the value is 1.

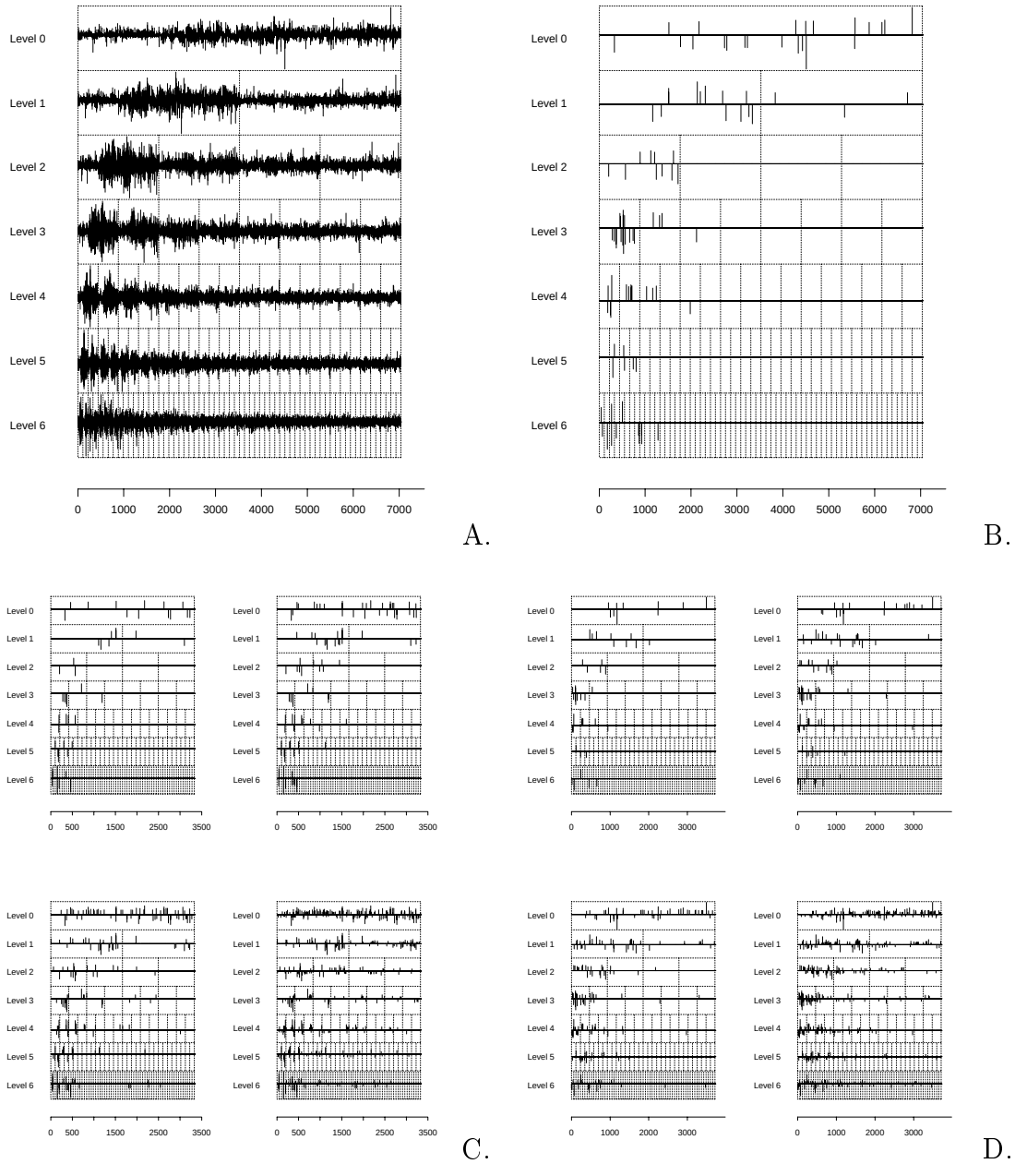


Figure 1: WP time-frequency tiling (A) and MP approximation with 100 atomic structures (B). MP progressive approximation power for 1st (C) and 2nd (D) sub-samples, with 50 (top left), 100 (top right), 200 (bottom left) and 500 atoms.

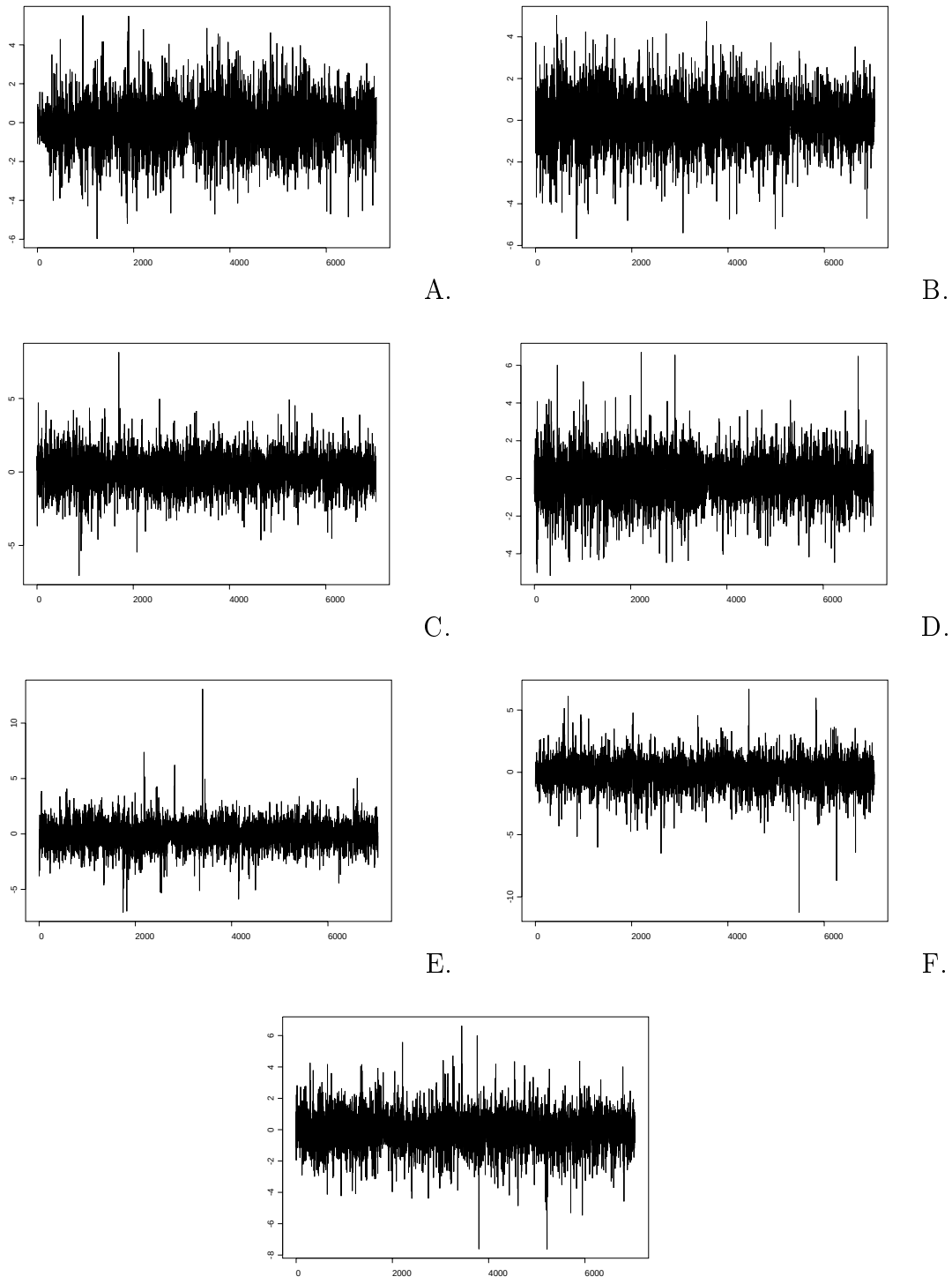


Figure 2: Extracted WP sources, from highest (A) to lowest (F) resolution level.

which might also be considered a wavelet periodogram. We have limited the analysis to WP, for the remaining experiments, due in particular to the interest we have in the wavelet system itself. The results we have obtained are particularly informative, particularly when relating the log-energy plot obtained of the WP coefficient sequences to that of the correspondent variance plot.

Note that a process is self-similar with Hurst parameter $H > 0$ if $X(st)$ and $s^H X(t)$ have identical distributions (see Figure 9). With long range dependence or asymptotic self-similarity $\alpha = 2H - 1$. The self similarity typical of financial series is here subject to the lens of the wavelet packet sequences, and it cannot be really stated exactly its presence due to a power law behavior that is not homogeneous across all scales, but only within a range of them.

The energy function refers to the second moment of the observed process at each resolution level, and thus represents a non-parametric unbiased estimator of the variance of the WP coefficients sequence. The power law structure of the function $E(WP_{j,k}^2)$ usually suggests that E_j is that of a self-similar or of a long range dependent process, due to a log-linear relation with slope $2H - 1$ or α, respectively.

A linearly decreasing behavior for the highest scales is followed by an upward step when resolution level 5 appears, and then a new jump occurs at level 6, addressing what already observed in Figure 3-A. This behavior is in part expected, being the scaling effects usually observed only in a limited range of scales, but might also depend on the presence of real multi-scaling regimes responsible for the non-linear signatures characterizing the energy plot.

The source energy-plots allow though an even more clear identification of regimes at various scales, since they are by construction almost independently informative, certainly to a greater degree compared to the WP coefficient sequences. Note that for both sequences, the range of scales 4-6 indicates the presence of regimes of variability, and possible bias over the last scale measure, which might also be due to a certain degree of latent periodicity in the series around the scale where the jump occurs (Gilbert, 2001).

When comparing the estimated spectrum and adopting a wavelet decomposition of the periodogram, for then reconstructing the spectrum, the application of a thresholding algorithm to the log periodogram $\Pi(w_k) = log \mid X(w_k) \mid$, where $X(w_k)$ is the discrete Fourier transform and $w_k = 2\pi k$ are the fundamental frequencies, one can see the effect of transforming the data (Figure 5 top plots) and removing noise from them (middle and bottom plots).

It appears that the wavelet spectral estimator adapts well to the general shape of a typical smoothed periodogram (which we have computed with triangular spectral windows of various widths) and smooths it out furtherly. No substantial boundary effects appear given the large fluctuations at the highest frequencies, which means that the wavelet thresholding[9] works like a variable bandwith smoother that results quite robust.

Figures 6 and 7 are auto- and cross-correlation functions computed for the highest two wavelet packet sequences and correspondent WP sources; one may note that the sources improve both the autocorrelation structure (as noted also in Figure 8, where an initial segment is analyzed) and the cross-correlation structure.

Figure 9 shows the density plots for the highest and lowest wavelet packet sequence and WP sources; the support range with higher density is substantially wider for the sources

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

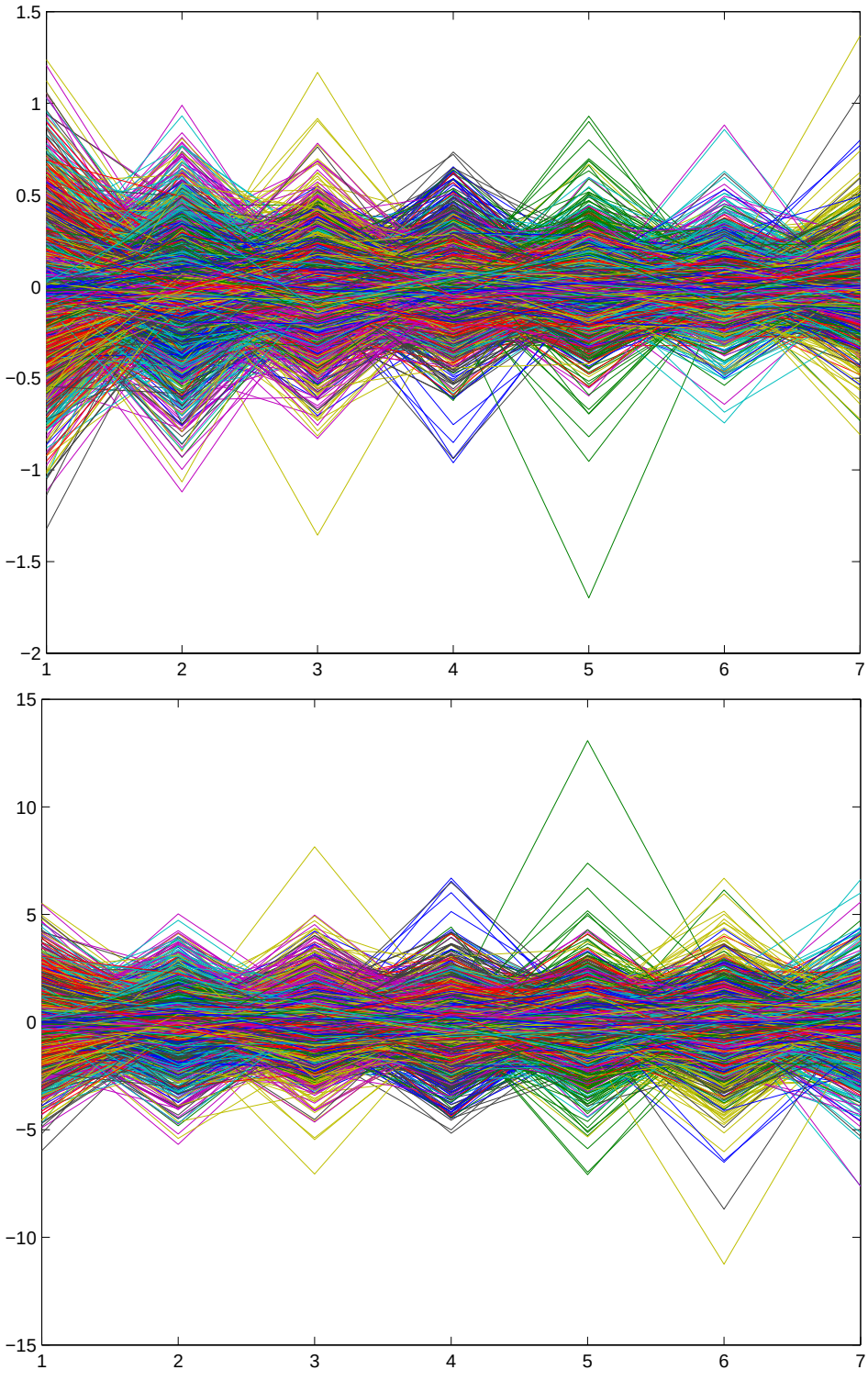


Figure 3: Resolution-wise variability for WP (A) and sources (B) sequences.

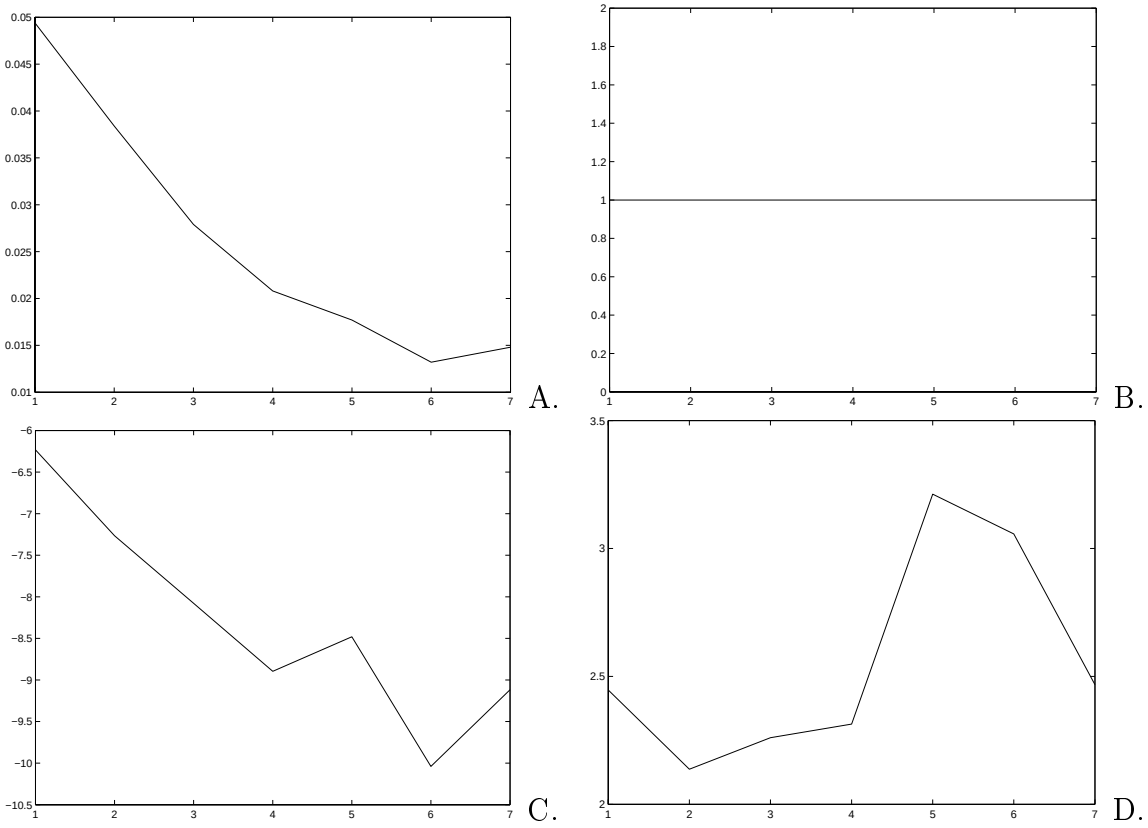


Figure 4: WP scale-wise variance for packet (A) and sources (B); log_2-energy plots for packets (C) and sources (D).

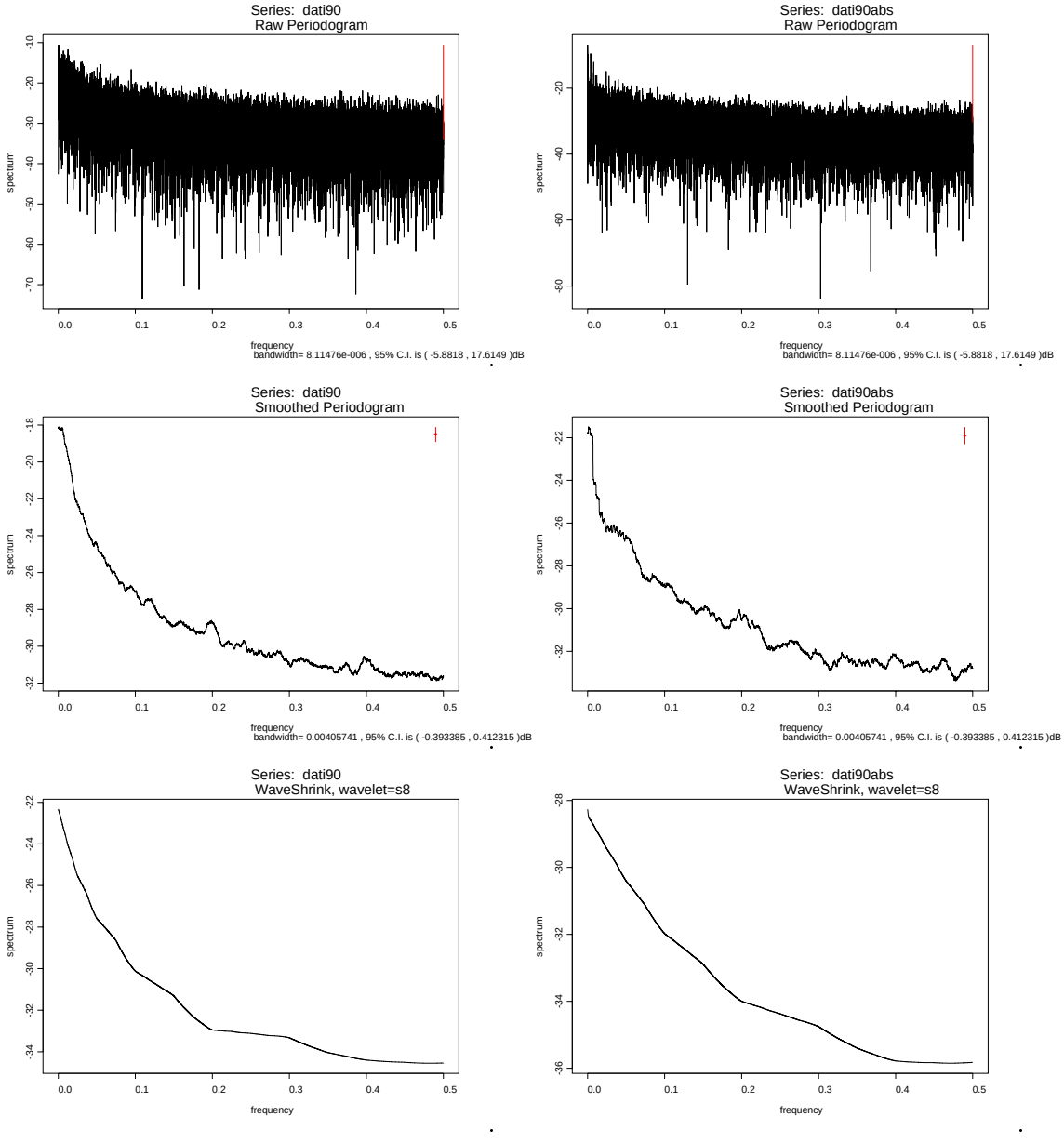


Figure 5: Top: raw periodogram. Middle: smoothed periodogram. Bottom: wavelet de-noised spectral estimate. Original 1m returns (left) and absolute values (right).

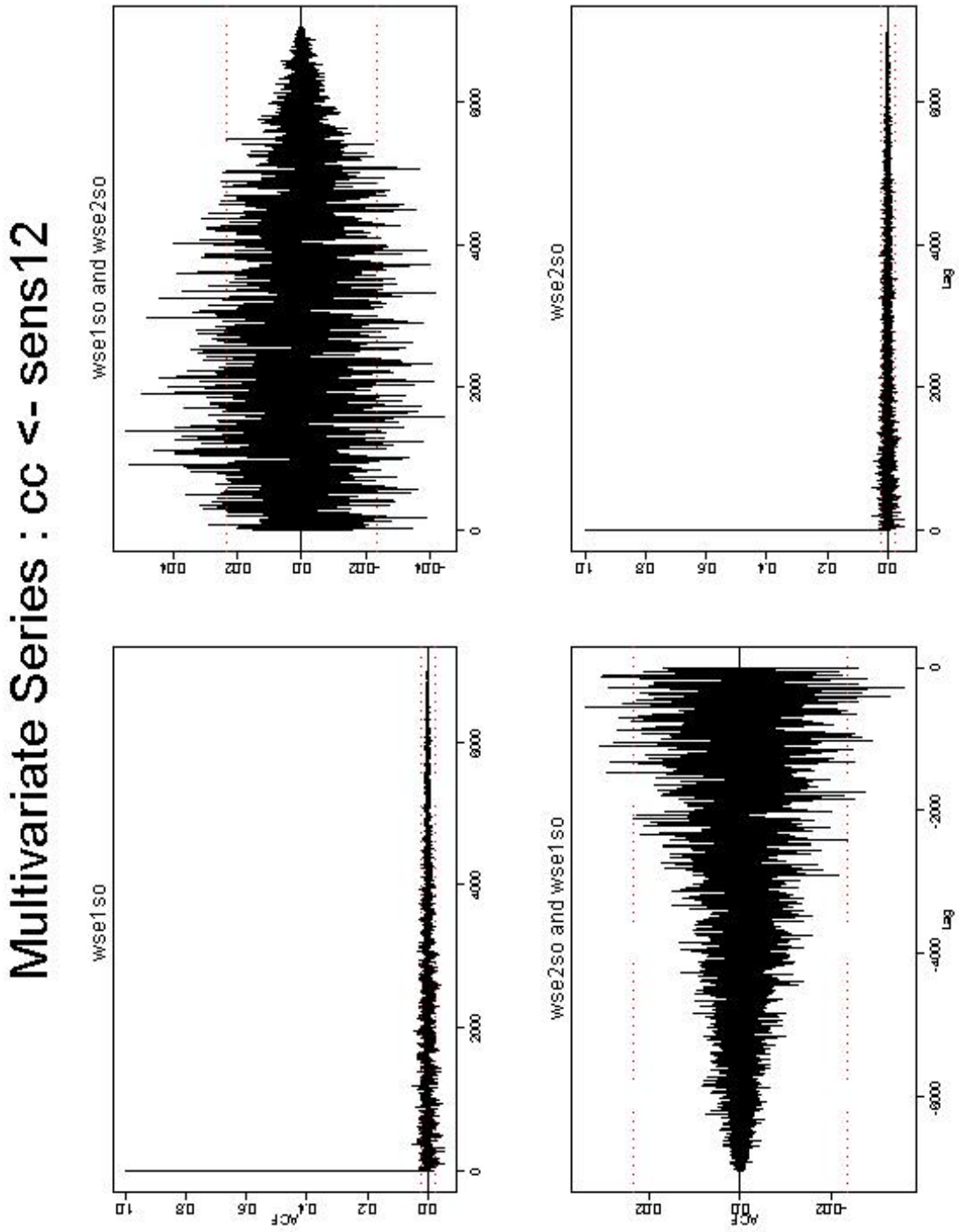


Figure 6: Auto- and Cross-correlation functions for first and second resolution WP sequences.

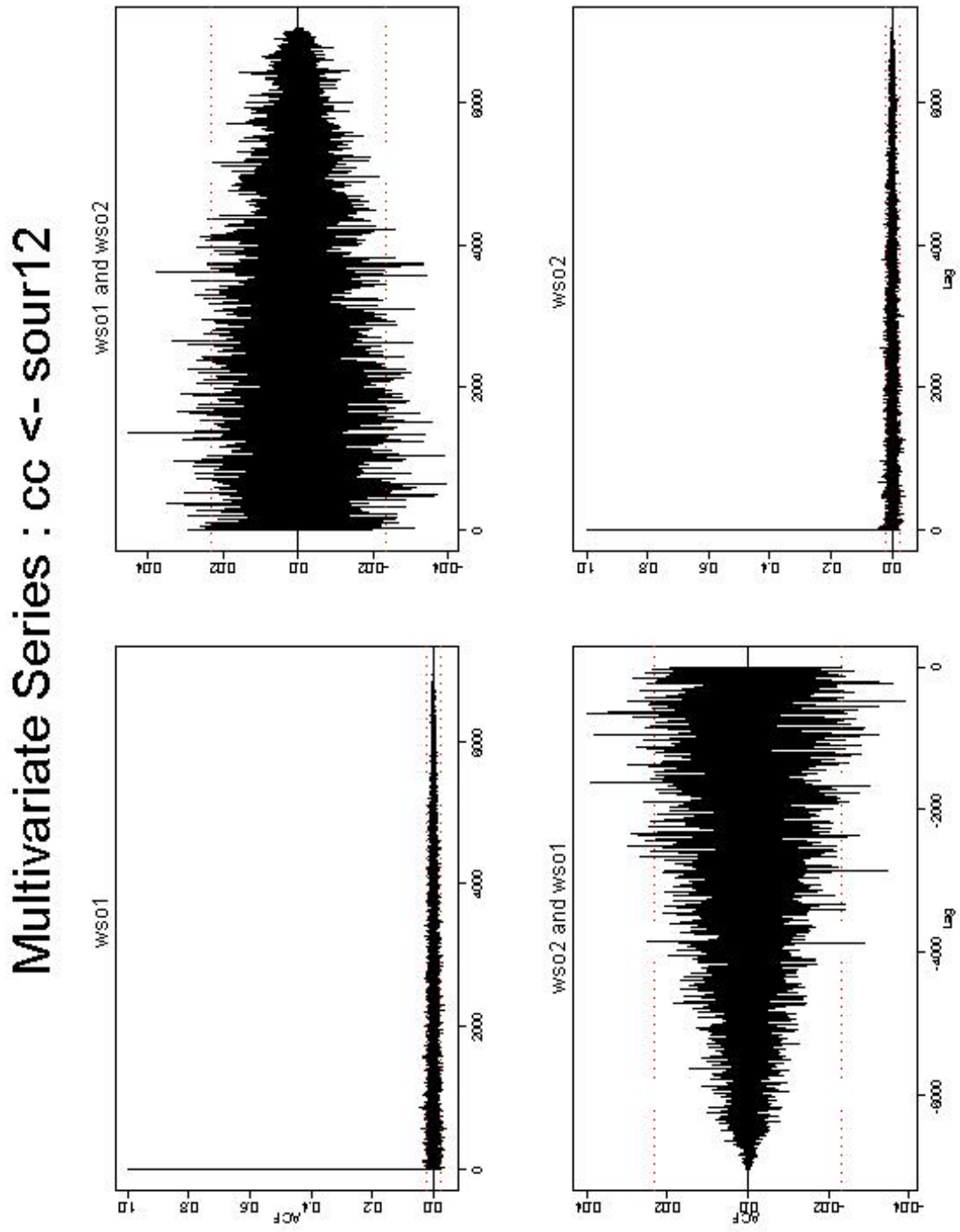
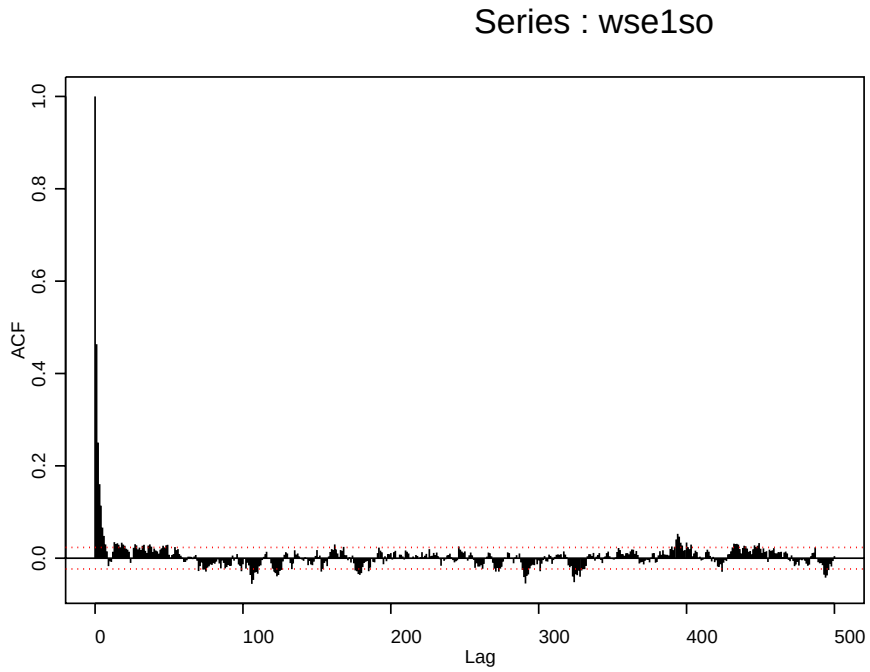
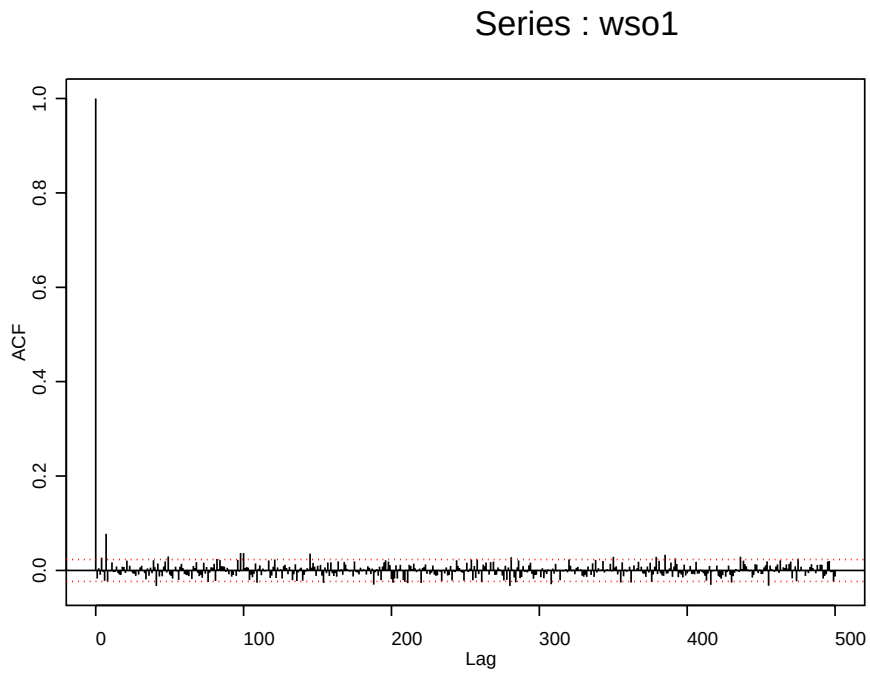


Figure 7: Auto- and Cross-correlation functions for first and second resolution WP sources.



A.



B.

Figure 8: ACF over a reduced sample for the finest scale WP sequence (A) and the first estimated source (B).

than the other values (see Figure 9, A-E and compare with C-G). The quantile plots in the right side of Figure 9 confirm what said, given the wider support range in B-F, related to WP sources, compared to D-H.

Figure 10 shows the relation of dependence between the WP sequences at different scales, compared to the same relations existing between the correspondent sources. One can notice that for the latter, a wider support region is covered in the scatter plot, suggesting that more independence has been achieved through the substantial dimensionality reduction of the problem. In other words, the reproducing kernel property which embeds the information and correlation core of the data in the projected space is successfully affected by the ICA procedure.

Then, Figure 11 shows scatter plots of pairs of wavelet packet and source sequences, this time combined and computed at the resolution levels. It shows that the structure of dependence within each scale and between the two different sequences changes toward a more independent relation. It does so to a slighter degree for the first scales compared to the across-scale relations seen before, due to the fact that the WP coefficients sequences are already yielding good decorrelation in the system. The last two levels show much more independence, due to a stronger influence of the other scales.

7. Conclusions

In this paper we have analyzed an high frequency stock index return series and have investigated a volatility model with risk and return dynamics linked through a flexible form approximating a non-linear relation. Mixture and cascade dynamics can be thought to characterize the model too, and we have found that the presence of regimes in the volatility structure might be justified in terms of scale-dependent behavior. Wavelet-like representations allow for multi-scale decompositions and thus become very useful instruments, particularly when combined with statistical decomposition techniques such as Independent Component Analysis. The learning task is then left to a greedy approximator, which in relatively few iterations can deliver near-optimal performance.

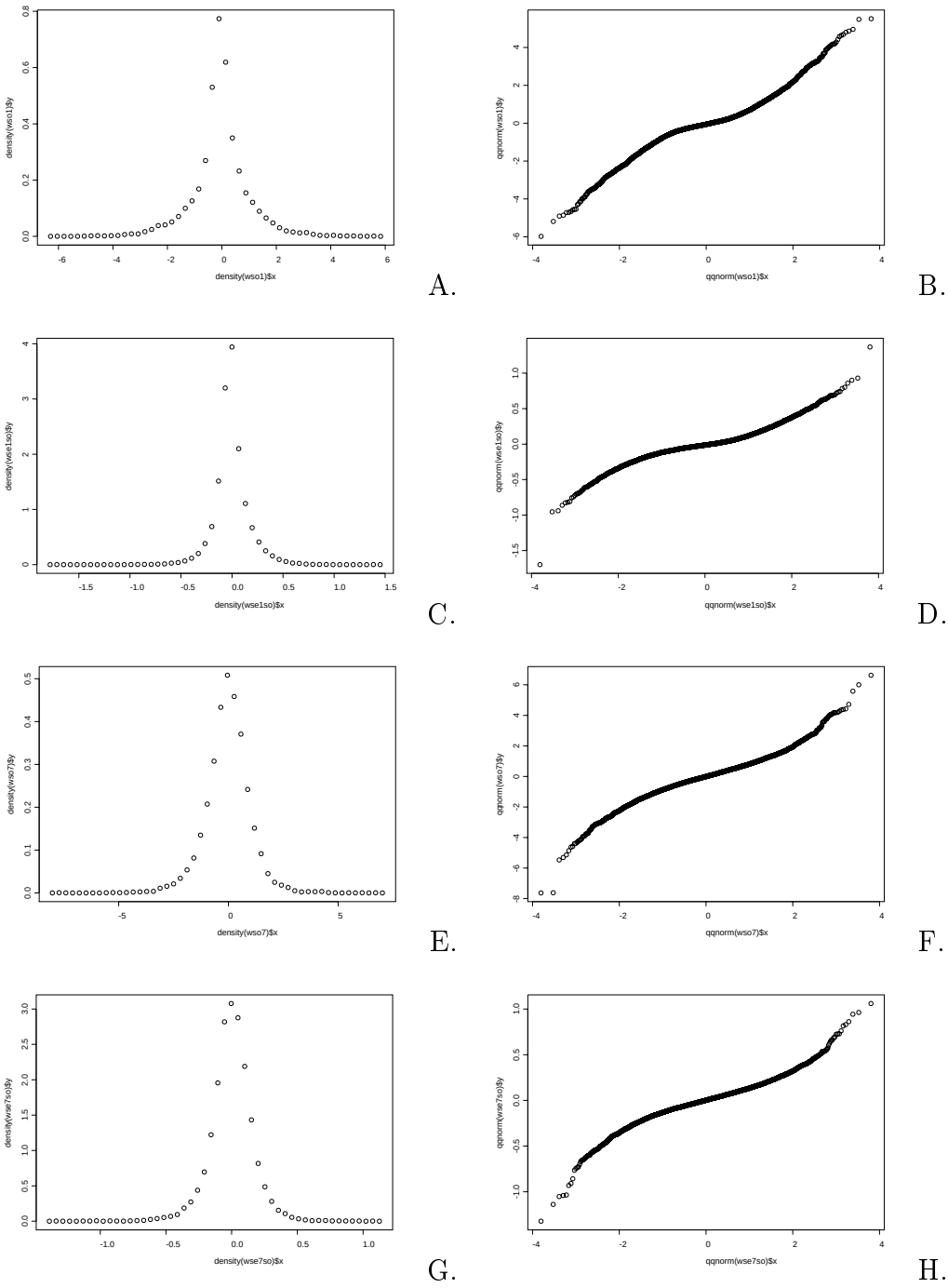


Figure 9: Density plots for WP highest/lowest resolution sequences (A/E) and source (C/E), and correspondent quantiles on left column.

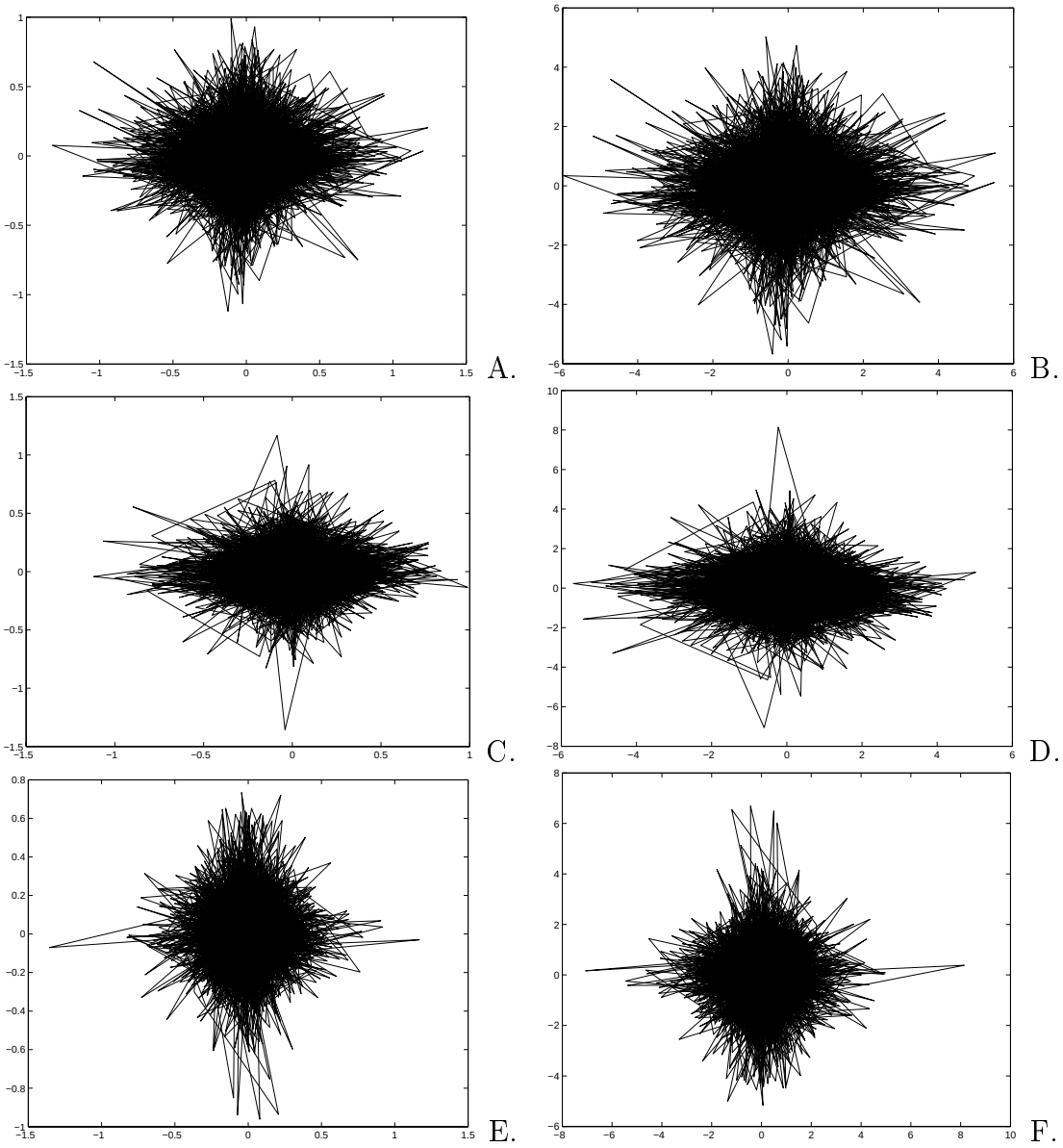


Figure 10: Some across-scale relations between WP sequences (left col.) and sources (right col.): levels 1-2 top, levels 2-3 middle, levels 3-4 bottom.

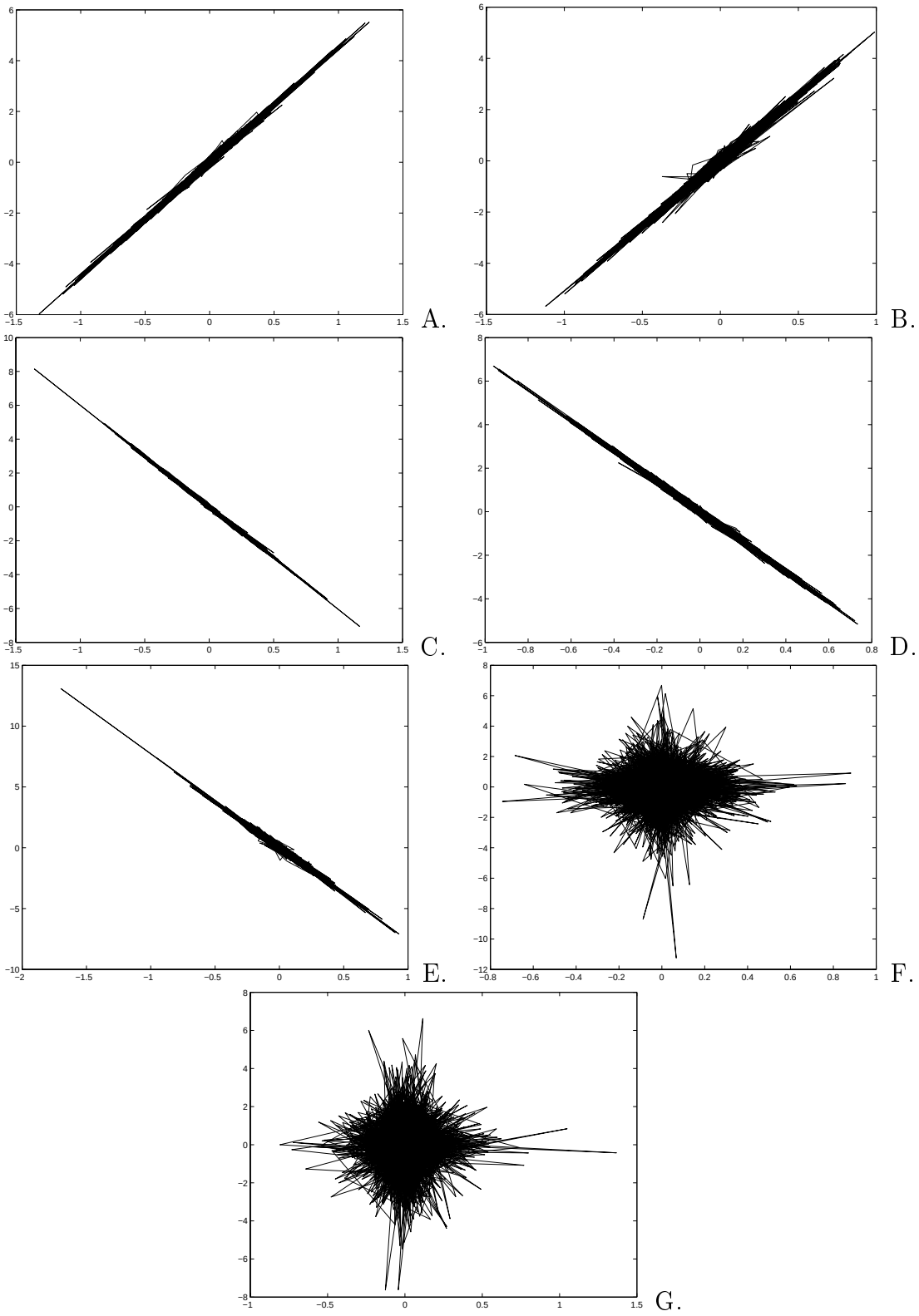


Figure 11: Within-scale relation between WP sequences and sources at all resolution levels (highest at A and lowest at G).

References

.

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

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

Andersen, T.A., (1994). Stochastic Autoregressive Volatility: a framework for volatility modeling. *Mathematical Finance*, 4, 75-102.

Andersen T. and Bollerslev T., (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39, 885-905.

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

Arneodo A., Bacry E. and Muzy J.F., (1998). Random cascades on wavelet dyadic trees. *Journal of Mathematical Physics*, 39(8), 4142-4164.

Aronszajn N., (1950). Theory of Reproducing Kernels. *Transactions American Mathematical Society*, 686, 337-404.

Backus D.K. and Gregory A.W., (1993). Theoretical relations between risk premiums and conditional variances. *Journal of Business and Economic Statistics*, 11, 177-185.

Barndorff-Nielsen O.E. and Shephard N., (2001). Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, B*, 63, 167-241.

Berkner K. and Wells R.O., (1998). A Correlation-Dependent Model for Denoising via Nonorthogonal Wavelet Transform, Technical Report, CML TR 98-07, Rice Un. (US) (1998).

Bollerslev T., (1986). A generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307-327.

Bruce A. and Gao H.V., (1994). *S+Wavelets*, StaSci Division, MathSoft Inc., Seattle.

Capobianco E., (2002a). Multiresolution Approximations for Volatility Processes. *Quantitative Finance*, 2, 91-110.

Capobianco E., (2002b). Hammerstein System Representation of Financial Volatility Processes. *European Physical Journal B*, 27(2). 201-212.

Capobianco E., (2002c). Independent Component Analysis and Resolution Pursuit with Wavelets and Cosine Packets. *Neurocomputing*, forthcoming.

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

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

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

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

Chou, R., Engle, R.F. and Kane, A., (1992). Measuring risk aversion from excess returns on a stock index. *Journal of Econometrics*. 52, 201-224.

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

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

Daubechies I., (1992). *Ten Lectures on wavelets*. SIAM, Philadelphia.

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

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

Donoho D., (2001). Sparse Components of Images and Optimal Atomic Decompositions. *Constructive Approximations*, 17, 353-382.

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

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

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

Durbin J. and Koopman S.J., (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *Journal of the Royal Statistical Society, B*, 62, 3-56.

Engle, R.F., (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50, 987-1008.

Engle, R.F., Lilien, D.M. and Robins, R.P., (1987). Estimating time varying risk premia in the term structure: the ARCH-M model. *Econometrica*, 55, 391-407.

Ghysels, E., Harvey A.C. and Renault, E., (1996). Stochastic Volatility. In: Maddala G.S., Rao C.R. (Eds.), *Handbook of Statistics*, vol. 14., Elsevier Science, Amsterdam, pp

118-191.

Gilbert A.C., (2001). Multiscale Analysis and Data Networks. *Applied and Computational Harmonic Analysis*, 10, 185-202.

Greblicki W., (2000). Continuous-Time Hammerstein System Identification. *IEEE Transaction on Automatic Control*, 45(6), 1232-1236.

Guegan D., (1990). Les modeles ARCH univaries. Working Paper n. 90-2, Universite Paris Nord.

Hasiewicz, Z., (2001). Non-parametric estimation of non-linearity in a cascade time-series by multiscale approximation. *Signal Processing*, 81, 791-807.

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

Johnstone I.M., Silverman B.W., 1997. Wavelet threshold estimators for data with correlated noise. *Journal Royal Statistical Society, B*, 59, 319-351.

Karatzas I. and Shreve E., (1988). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.

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

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

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

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

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

Ralston J.C., Zoubir A.M. and Boashash B., (1997). Identification of a Class of Nonlinear Systems under Stationary Non-Gaussian Excitation. *IEEE Transactions on Signal Processing*, 45(3), 719-735.

Taylor, S., (1986). *Modeling Financial Time Series*. John Wiley, Chichester.

Taylor, S., (1994). Modeling Stochastic Volatility: a review and comparative studies. *Mathematical Finance*, 4, 183-204.

Temlyakov, V.N., (2000). Weak Greedy Algorithms. *Advances in Computational Mathematics*, 12(2,3)213-227.

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

Weiss, A.A., (1986). ARCH and bilinear time series models: comparison and combination. *Journal of Business and Economic Statistics*, 4, 59-70.