Wavelet Transform in Similarity Paradigm II

Z.R. Struzik, A. Siebes

Information Systems (INS)

INS-R9815 December 1998

Report INS-R9815
ISSN 1386-3681

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

Wavelet Transform in Similarity Paradigm II

Zbigniew R. Struzik, Arno Siebes

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
email: Zbigniew.Struzik@cwi.nl

ABSTRACT

For the majority of data mining applications, there are no models of data which would facilitate the tasks of comparing records of time series, thus leaving one with 'noise' as the only description. We propose a generic approach to comparing noise time series using the largest deviations from consistent statistical behaviour.
For this purpose we use a powerful framework based on wavelet decomposition, which allows filtering polynomial bias, while capturing the essential singular behaviour. In particular we are able to reveal scale-wise ranking of singular events including their scale-free characteristic: the Hölder exponent.
We use such characteristics to design a compact representation of the time series suitable for direct comparison, e.g. evaluation of the correlation product. We demonstrate that the distance between such representations closely corresponds to the subjective feeling of similarity between the time series. In order to test the validity of subjective criteria, we test the records of currency exchanges, finding convincing levels of (local) correlation.

1991 Mathematics Subject Classification: 28A80, 65U05, 68T10, 68P10
1991 Computing Reviews Classification System: H.1, I.5, J.m, J.2, E.2
Keywords and Phrases: wavelet transform, Hölder exponent, time series correlation, similarity measure.
Note: This work has been carried out under the Impact project.

1. Introduction

The issue of quantitative similarity estimation between time series in data mining applications seemingly suffers from a serious internal inconsistency; on the one hand one wants the similarity to be independent of a large class of linear transformations like (amplitude, time) rescaling, addition of linear trend or constant bias. This is understandable since most such operations affect the parameter values of commonly used estimators (e.g. power spectrum), or destroy any stationarity potentially present in the time series making estimation impossible. At the same time, the subjective, qualitative judgment of similarity (by humans) is based precisely on non-stationary behaviour; rapid transients marking beginnings of trends, extreme fluctuations and generally speaking, strong but rare events.

Generically, time series data for which there is no model are treated as noise following a certain distribution (or alternatively a power spectrum, another global measure). Comparing such time series is an awkward task - in practice therefore one resigns oneself to matching these global measures, distributions or spectra. Naturally, the discrimination power of such tests is restricted to the universality class of the representation applied - the spectral test will not distinguish between different processes if they have the same spectrum but differ in distribution. By the same argument, processes with the same distribution may have different spectra but will not be distinguished if only the distributions are tested.

Still, the global statistical characterisation of this type will usually be inferior to characterisation by humans when determination of similarity (correlation or matching) between time series is required. Even without a model, the human observer is capable of identifying and localising rapid transients, trends and fluctuations in the data which best characterise the time series in question. Even with the same distribution and power spectrum, the same process can have two realisations which differ in local detail. In addition to this, such local fluctuations characteristic for single realisations usually

make statistical estimations difficult and result in unreliable estimates. In particular, it is common knowledge that the evaluation of data distributions from short data sets is an awkward task, resulting in unreliable estimates. The reason for this is limited statistics, in which local fluctuations of the data override consistent statistical behaviour. However, what is of great disadvantage from the statistical point of view can be of advantage in another context. In this report, we propose a method of characterising the time series which relies on such deviations from the consistent statistical behaviour as caused by the non-stationary behaviour of the data. We will show how large local fluctuations in relatively short data sets carry the relevant information about the transient 'shape' of the time series. In particular we can then make use of them in order to provide a very compact set of characteristics of the time series useful for correlation or matching purposes.

But what if the time series data in our application is long enough to result in good statistical estimates? The way to go is, of course, to reduce the data length in order to increase the influence of large local fluctuations! What sounds unreasonable, is perfectly admissible and technically possible, by the operation of coarse graining the data using so-called *wavelet* filters, in the Wavelet Transformation scheme.

In the previous work [1], we have used this recently introduced tool - the Wavelet Transformation (WT), capable of characterising the time series, independently of translation and polynomial bias but also of scaling and normalisation. We have shown how to use the WT based representation to define similarity measures in two extreme formulations: the global - statistical similarity and the localised, detail oriented case. In both cases, we have used the representation capturing scale-wise hierarchy of singular events in the time series, the Wavelet Transform Modulus Maxima or Bifurcation representations. We have also shown that the Wavelet Transform method of scale-wise decomposing time series data can be successfully used to characterise such singular behaviour by means of global and local scaling exponents.

In analysing these features we did not, however, pay attention to their relative scale localisation, or the frequency of occurrence of the certain value of the scaling exponent estimate. In other words we treated the feature space of the WT of the data on an equal, flat basis, without any attempt to rank the derived features. As noted above, it is often the case that it is not the statistical bulk behaviour but some rare, extreme events, or alternatively, the strong, local fluctuations of the time series, which determine the interestingness of the time series. Still, if treated on an equal basis with all information in the time series, they would be overshadowed by the 'noise' of consistent behaviour.

In this report we will demonstrate how the Wavelet Transform method of scale-wise decomposing time series data provides a natural method to obtain scale-wise ranking of events in the time series. In addition to this, by evaluating both the local scaling estimates and the spectral density of singular behaviour in the time-series, we will be able locally to indicate rare events in time-series. These will next be used for the purpose of (locally) correlating time-series using large or rare events.

In section 2, we will give the heuristic motivation for the methods to be described. It will be argued that the rare events and large fluctuations will both provide the information relevant to distinguishing between statistically indistinguishable time series. In section 3, we will focus on the relevant aspects of the wavelet transformation, in particular the ability to characterise scale free behaviour of characteristic events in time series, like 'crash' singularities. The link of such singularities with the non-stationary behaviour of time series will be postulated, and together with the hierarchical scale-wise decomposition provided by the wavelet transform, it will enable us to select the interesting large scale features.

In section 4, we will introduce a technical model enabling us to estimate the scale-free characteristic (the effective Hölder exponent) for the thus selected large scale events, for real life time-series, i.e. in the case of dense singularities. In section 5, we will discuss the h-representation of time series, utilising the large scale characteristics with exponents properly estimated. The issues of distance metric in the representation and that of correlation between the representations will be addressed. This is followed by the test case of correlating examples of currency exchange rates in section 6. Section 7 closes the report with conclusions and suggestions for future developments.

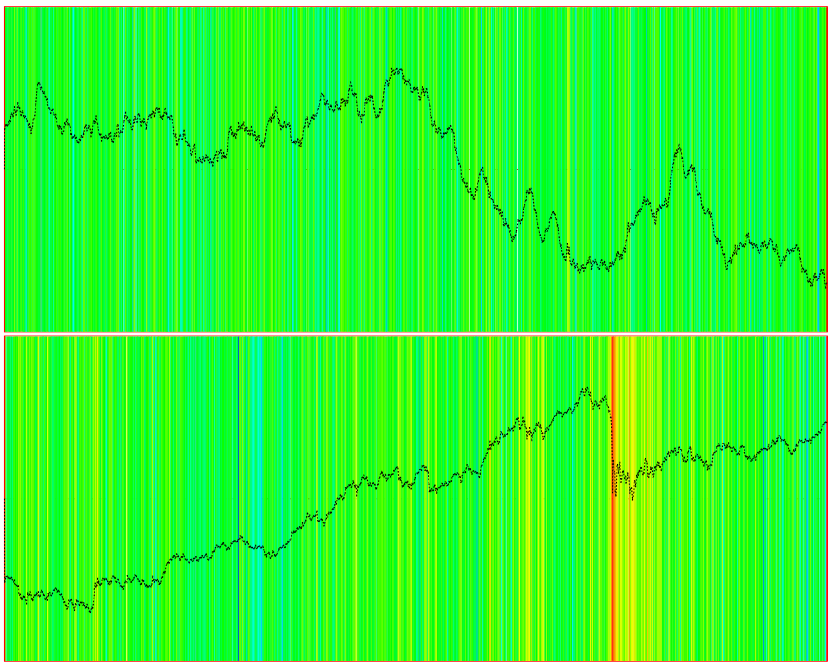

Figure 1: Two time series with local Hurst exponent indicated in colour. Both time series have the same global Hurst exponent but differ in distribution. The relevant detail information is contained in rare events in the second time series. The time series above has a Gaussian distribution of jumps, while the time series below has a *fat* tailed distribution.

2. Rare Events versus Large Fluctuations

Suppose, by virtue of example, that we want to correlate several stock indexes in order to find their degree of dependence on one another. Direct implementation of a standard correlation product would probably not give very exciting results. Generally, the stocks would have different values and could have different sampling rates. Some might respond better to some market stimuli and have a higher linear trend than others. Plus, they would, of course, have a substantial constant bias, as the result of the history of the index.

In addition to these problems, the bulk of financial data, including indices consists mainly of nearly brown noise (Hurst exponent; $H \sim 0.5$) [2]. Even if we manage, despite of the non-stationary biases just mentioned, correctly to evaluate its distribution, this will be entirely useless for the purpose of discriminating (or correlating) one such time series from another. As two realisations of the same statistical process, they will simply follow the same distribution and will, therefore, be indistinguishable from a statistical point of view.

Still, the common sense of trading would mark as correlated the indices which responded in a similar way to the same *largest* stimuli, neglecting the small scale 'noise' on the data. Not coincidentally, such large singular jumps define the difference between pure brown noise and financial records - the probability of strong rare events is considerable and higher than for the Gaussian distribution, see figure 1, and Ref. [3]. It is thus quite evident that a representation effectively filtering out the low level noise and retaining the strongest rare events would be required for a reliable and efficient correlation of financial indexes.

In figure 1, we show two example time series. Both consist of nearly uncorrelated noise with the same Hurst exponent, close to that of the Brownian walk ($H_B = 0.5$). Using the method which is described in this report, we have indicated the local version of the Hurst exponent, the Hölder

exponent, locally along the time series. (Note: Just as the Hurst exponent can be considered a global roughness indicator, the Hölder exponent can be loosely associated with the feeling of local roughness or regularity of the time series.) This is done with colour ranging from blue for minimum to red for maximum with green centered at the mean value.[1] As an immediate implication of such a procedure comes the observation that the spectral density of the lower time-series is much richer than that of the time-series above. Indeed, we can verify that the even though both time series have the same global Hurst exponent, they differ in distribution (of singular events). The time series above has a Gaussian distribution of jumps, while the time series below has fat tailed distribution. It is a record [1984-1988] for the S&P index [3].

Thus, the relevant detail information in the second time-series is contained in rare events marked with less frequent colour. Note that these rare events are not necessarily the largest events in the sense of the absolute index value, but they do closely correspond to the most smooth (blue/white) and the most singular (red/black) events.

At the same time the first time series remains essentially monochromatic. While for some applications, stating this fact may be sufficient, it is evident that the particular realisation of the process involved has resulted in large events - the fluctuations of the size close to that of the sample length. These only weakly distort the almost perfectly monochromatic/narrow band of singularities in the time series. The small colour noise on the monochromatic background for such a high resolution analysis is mainly caused by deviations at the very small resolution considered. By means of analysing the colour distribution for smaller and smaller data lengths, obtained with the WT filters, we will however be able to give characterising power to such fluctuations at large scales. Even though they follow the same statistical rule of the Gaussian process, for some applications their location may well be as crucial as the presence and the location of rare events in the lower time series in figure 1.

3. Continuous Wavelet Transform and its Maxima Used to Reveal the Structure of the Time Series

As already mentioned above, the recently introduced Wavelet Transform (WT), see e.g. Ref. [4], provides a way of analysing local behaviour of functions. In this, it fundamentally differs from global transforms like the Fourier Transform. In addition to locality, it possesses the often very desirable ability of filtering the polynomial behaviour to some predefined degree. Therefore, correct characterisation of time series is possible, in particular in the presence of *non-stationarities* like global or local trends or biases. Last described but certainly not least for our purpose, one of the main aspects of the WT which is of great advantage is the ability to reveal the *hierarchy* of (singular) features including the scaling behaviour - the so-called *scale-free* behaviour. We will omit the formal aspects of the wavelet transform from this report, referring the reader to more complete specialised treatment [4, 5]. We will, however, highlight the key concepts mentioned above, also describing them in detail adequate to our problem.

Conceptually, the wavelet transform is a convolution product of the time series with the scaled and translated kernel - the wavelet $\psi(x)$, usually a $n-th$ derivative of a smoothing kernel $\theta(x)$. Usually, in the absence of other criteria, the preferred choice is the kernel well localised both in frequency and position. In this report, we chose the Gaussian $\theta(x) = \exp(-x^2/2)$ as the smoothing kernel, which has optimal localisation in both domains.

The scaling and translation actions are performed by two parameters; the scale parameter s 'adapts' the width of the wavelet kernel to the *microscopic resolution* required, thus changing its frequency contents, and the location of the analysing wavelet is determined by the parameter b:

$$Wf(s,b) = \frac{1}{s} \int_{-\infty}^{\infty} dx\ f(x)\ \psi(\frac{x-b}{s})\ , \tag{3.1}$$

where $s, b \in \mathbf{R}$ and $s > 0$ for the continuous version (CWT).

[1] In b/w version, gray-level coding is used ranging from white for minimum to black for maximum.

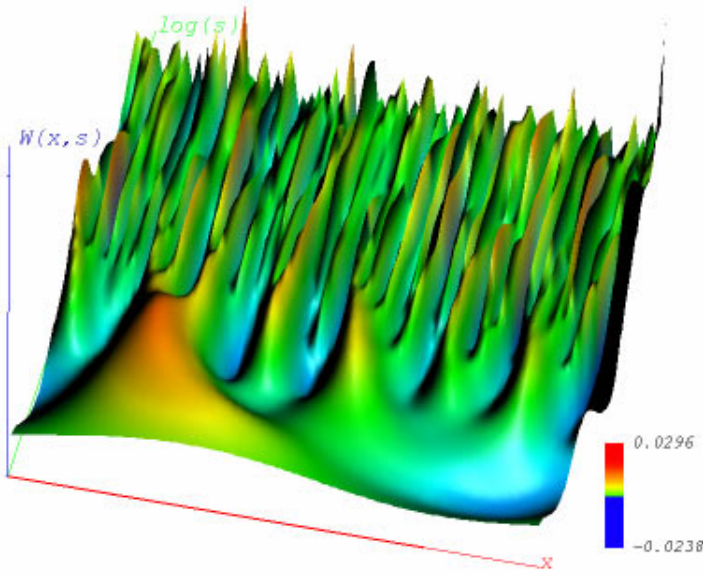

Figure 2: Continuous Wavelet Transform representation of the random walk (Brownian process) time series like that in figure 1 top. The wavelet used is the Mexican hat - the second derivative of the Gaussian kernel. The coordinate axis are: position x, scale in logarithm $\log(s)$, and the value of the transform $WT(s,b)$.

In figure 2 we show the wavelet transform of a random walk sample decomposed with the Mexican hat wavelet - the second derivative of the Gaussian kernel. From the definition, the transform retains the entire temporal locality properties - the position axis is in the forefront of the 3D plot. (In this report, we will mainly refer to position denoted x or b, meaning position in the time axis.) The standard way of presenting the CWT is using the logarithmic scale, therefore the scale axis pointing 'in depth' of the plot is log(s). The third vertical axis denotes the magnitude of the transform $W(s,b)$.

The 3D plot shows how the wavelet transform reveals more and more detail while going towards smaller scales, i.e. towards smaller $\log(s)$ values. Therefore, the wavelet transform is sometimes referred to as the 'mathematical microscope', due to its ability to focus on weak transients and singularities in the time series. The wavelet used determines the optics of the microscope; its magnification varies with the scale factor s.

3.1 Accessing Singular Behaviour with the Wavelet Transformation

Quite frequently it is the singularities, the rapid changes, discontinuities and frequency transients, and not the smooth, regular behaviour which are interesting in the time series. Let us, therefore, demonstrate the wavelet's excellent suitability to address singular aspects of the analysed time series in a *local* fashion. The singularity strength is often characterised by the so-called Hölder exponent - if we represent the function f through its Taylor expansion around $x = x_0$:

$$f(x)_{x_0} = c_0 + c_1(x - x_0) + \cdots + c_n(x - x_0)^n + C|x - x_0|^{h(x_0)} \ . \tag{3.2}$$

The exponent $h(x_0)$ is termed the Hölder exponent of the Hölder singularity at x_0. It follows directly that if $h(x_0)$ is equal to a positive integer n, the function f is n times continuously differentiable in x_0. Alternatively, if $n < h(x_0) < n + 1$, the function f is continuous and singular in x_0. In this case, f is n times differentiable, but its n^{th} derivative is singular in x_0 and the exponent h characterises

this singularity. The exponent h, therefore, gives an indication of how regular the function f is in x_0, that is the higher the h, the more regular the function f.

The wavelet transform of the function f in $x = x_0$ with the wavelet of at least n vanishing moments, i.e. orthogonal to polynomials up to (maximum possible) degree n:

$$\int_{-\infty}^{+\infty} x^m \, \psi(x) \, dx = 0 \qquad \forall m, \; 0 \leq m < n \; ,$$

reduces to

$$W^{(n)} f(s, x_0) = \frac{1}{s} \int C|x - x_0|^{h(x_0)} \; \psi(\frac{x - x_0}{s}) \; dx = C|s|^{h(x_0)} \int |x'|^{h(x_0)} \; \psi(x') \; dx' \; .$$

Therefore, we have the following scale-wise proportionality of the wavelet transform of the (Hölder) singularity $n \leq h \leq n + 1$, with the wavelet with n vanishing moments:

$$W^{(n)} f(s, x_0) \sim |s|^{h(x_0)} \; .$$

Thus the continuous wavelet transform can be used for detecting and representing the Hölder singularities in the time series even if masked by the polynomial bias. Note: we will restrict the scope of this report to Hölder singularities, thus not taking into consideration the so-called oscillating singularities requiring two exponents [6]. The range of influence of a singularity on the wavelet transform is limited to the so-called *cone of influence*. It can be characterised by the standard deviation σ of the wavelet used and, therefore, increases linearly with the scale: $(x_0 - s) \leq \sigma$. In figure 3, we show the wavelet transform of the Dirac pulse with obtained with the Mexican hat wavelet. The range of influence of the Dirac pulse on the coefficients of the transform is by the above definition restricted to the 2D position-scale 'cone' originating at x_0 - the location of the Dirac pulse itself, and spreads within the bounds of two straight lines marking the standard deviation σ of the wavelet used. Note, that unlike in all other plots we used linear scale s here.

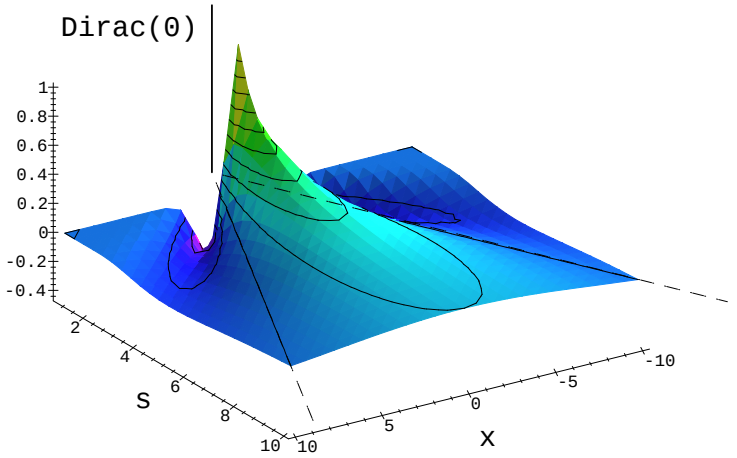

Figure 3: CWT representation of the single Dirac delta $D(x_0)$ located at $x_0 = 0$. It shows the so-called 'cone of influence' which in this case takes the form of a triangle pointed at x_0 and bounded by $(x_0 - s) \leq \sigma$. Due to the linear scale used, the local maximum line centered at $x_0 = 0$ follows $1/s$ increase, which in log-log scale gives -1 slope. The wavelet used is the Mexican hat.

It is not necessary to evaluate an entire cone of influence in order to characterise its related singularity. In fact any line converging to the singularity at hand within the range of the cone of influence can be used for this purpose [7]. In particular, this can be the line where the wavelet transform reaches local maximum (with respect to position coordinate). Connecting such maxima within the continuous wavelet transform 'landscape' gives rise to the so-called maxima lines. It turns out that restricting oneself to the collection of such maxima provides a particularly useful representation of the entire CWT. The fact that these lines generally converge to singular points in the signal is of course one of the important properties of this representation. In the following subsection we will describe this and other advantages of the representation using the maxima lines, the so-called Wavelet Transform Modulus Maxima (WTMM) representation.[2]

3.2 Wavelet Transform Modulus Maxima Representation

The continuous wavelet transform described in Eq. 3.1 is an extremely redundant representation, much too costly for most practical applications. This is the reason why other, less redundant representations, are frequently used. Of course, in going from high redundancy to low redundancy (or even orthogonality), certain (additional) design criteria are necessary. For our purpose of comparison of the local features of time series, one critical requirement is the translation shift invariance of the representation; nothing other than the boundary coefficients of the representation should change, if the time series is translated by some Δx.

A useful representation satisfying this requirement and of much less redundancy than the CWT is the Wavelet Transform Modulus Maxima (WTMM) representation, introduced by Mallat [8]. In addition to translation invariance, it also possesses the ability to characterise fully local singular behaviour of time series as illustrated in the previous subsection.

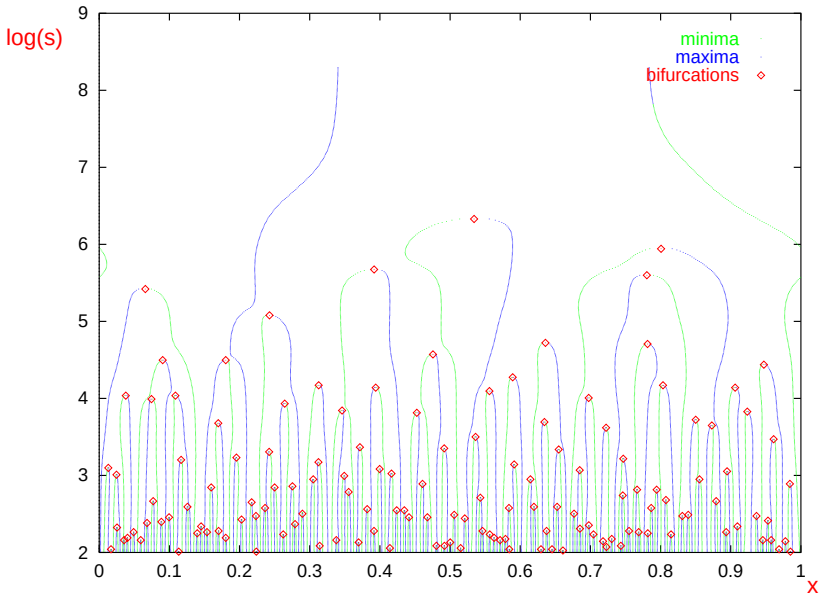

Figure 4: WTMM representation of the time series and the bifurcations of the WTMM tree. Mexican hat wavelet.

Both the aforementioned properties of the maxima lines representation make it particularly useful for our purpose. The WTMM is derived from the CWT representation by extracting lines of local maxima with respect to position/time (of the modulus) of the wavelet transform.

[2]Here we use 'modulus' and WTMM for historical consistency reasons. Also, within this report we limit our vocabulaly to the use of 'maxima', which will simply mean positive and negative maxima, (i.e. minima) of the CWT.

Since the necessary requirement for the maximum is zero of the derivative of the WT with respect to the position coordinate x, this can be used for the definition (and for the actual computation) of the maxima (minima) along scale:

$$\begin{cases} & \frac{d(Wf)(s,x)}{dx} = 0 & \text{and} \\ \text{either} & \frac{d^2(Wf)(s,x)}{dx^2} < 0 & \text{for maximum} \\ \text{or} & \frac{d^2(Wf)(s,x)}{dx^2} > 0 & \text{for minimum}\,. \end{cases} \tag{3.3}$$

An additional condition for zero of the second derivative identifies the beginning of the maximum (minimum) line, in the point of *bifurcation* or the so-called *top point*:

$$\begin{cases} \frac{d(Wf)(s,x)}{dx} = 0 \\ \frac{d^2(Wf)(s,x)}{dx^2} = 0\,. \end{cases} \tag{3.4}$$

The scale coordinate of the top point of a maximum line will be accordingly called *top-scale*. An example WTMM tree is shown in figure 4, together with the high-lighted bifurcations of the maxima lines [9].

While, as indicated in the previous subsection, the Hölder exponent of the singularity can be evaluated from the entire cone of influence, it is much more convenient to consider the maximum of the Wavelet Transform only. It can be shown that such a maximum converges to the singularity and that it can be used for the evaluation of the Hölder exponent of the singularity.

Let us consider the following set of examples, see figure 5 left; a single Dirac pulse at $D(1024)$, the saw tooth consisting of an integrated Heaviside step function at $I(2048)$, and the Heaviside step function for $S(3072^+)$, where $+$ denotes right-handed limit. The Hölder exponent of a Dirac pulse is -1, and each step of integration results in an increase of this exponent by 1. We, therefore, have $h = 0$ for the right sided step function $S(3072^+)$ and $h = 1$ for the integrated step $I(2048)$.

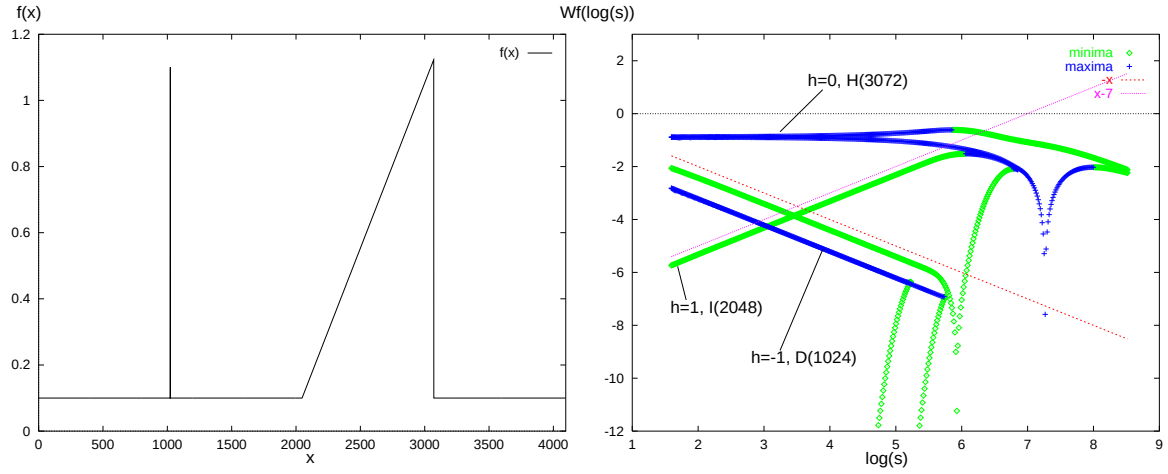

Figure 5: Left: The test signal consisting of the Dirac pulse $D(1024)$, the change in slope - integrated Heaviside step I(2048), and the Heaviside step H(3072). Right: The log-log plot of the maxima, together with their respective logarithmic derivative, corresponding to all three singularities: $D(1024)$, I(2048) and H(3072). Lines of theoretical slope are also indicated; these are $-x$ for $D(1024)$, x for I(2048) and a constant for H(3072). The wavelet used is the Mexican hat. Normalisation $1/s$.

For Hölder singularities, the process of integration and differentiation adds and subtracts one from the exponent. This can be also verified in the results obtained from the scaling of the maxima lines.

We obtain the (logarithmic) slopes of the maxima values very closely following the correct values of these exponents, see figure 5 right. This, of course, suggests the possibility of the estimation of the Hölder exponent of (Hölder) singularities from the slope of the maxima lines approaching these singularities. An important limitation is, however, the requirement for the singularities to be *isolated* for this procedure to work. Note that the scaling of the maxima lines becomes stable in the log-log plot in figure 5 right, only below some critical scale s_{crit}, below which the singularities effectively become isolated for the analysing wavelet. Indeed, the distance between the singular features in the test time-series in figure 5 left, equals 1024, which is in the order of three standard deviations of the analysing wavelet at $(\log(s_{crit}) = 5.83 = \log(1024/3)$.

3.3 Some Considerations on the WT Representation of Non-stationarities

As demonstrated above, the wavelet has to be orthogonal to polynomials up to a certain degree n in order to access the singularity exponent h by filtering out the polynomial bias. This operation of filtering the polynomial behaviour is nothing other than differentiating the time series to the degree n, the number of vanishing moments of the wavelet. This is evident from the fact that the Wavelet Transformation commutes (up to factor $-s$) with the operation of differentiation:

$$Wf^{(n)}(s,b) = \frac{1}{s}\int f(x)\ \psi^{(n)}(\frac{x-b}{s})\ dx = \frac{1}{s}\int f(x)\ (-s)^{(n)}\ \frac{d^n}{db^n}\theta(\frac{x-b}{s})\ dx = \tag{3.5}$$

$$= (-s)^{(n)}\ \frac{d^n}{db^n}\left(\int f(x)\ \theta(\frac{x-b}{s})\ dx\right) = (-s)^{(n)}\ D^{(n)}_{(\theta(s))}f(x) \tag{3.6}$$

Therefore, using wavelets with n vanishing moments, one can perform a stable derivation of the n-th order - one can obtain a smoothed derivative $D^n_{(\theta(s))}$ of the time series at the given scale s. The degree of derivation can be controlled with n, the number of vanishing moments. For $n = 1$, i.e. for the analysing wavelet orthogonal to constants, the first derivative of a smoothing function, we obtain the representation corresponding to the first derivative of the function, the local slopes of the input time series.

$$D_{(\theta(s))}f(x) = -\frac{1}{s^2}\int f(x)\ \psi(\frac{x-b}{s})\ dx \tag{3.7}$$

Note that the WTMM representation makes use of the maximum values of the same convolution product, compare Eq. 3.1, but with the normalisation factor set to $1/s$. The maxima lines are therefore proportional, locally in position and scale, to the strongest values of the first derivative of the analysed time series, smoothed with the smoothing kernel of the width proportional to $1/s$.

Let us summarize the above observations in more intuitive terms; at a particular scale of analysis, the Wavelet Transform gives the local derivative insensitive to the polynomial trend, global or large in relation to the working scale. This is how the effect of orthogonality to polynomials is achieved. At the same time, complete information about the trend is still preserved in the WT through the related singularities, in particular those marking the beginning and end of the trend. It allows among other things the reconstruction of the trend. This property of the Wavelet Transform makes it possible to compare time series with non-stationarities. The information about them becomes 'compressed' into the maxima, in a (semi)orthogonal fashion. Therefore, computing similarity measures with the WTMM or related representations, which we will use in the following, does not 'blow up' the correlation product due to non-stationarities, and allows the comparison of the features of the time series.

Let us show an example of a random walk sample with an artificially added step, see figure 6. The wavelet transform is obtained with the first derivative of the Gaussian smoothing kernel - the maxima representation shows two strong maxima starting from the largest scale - the highest resolution available. Both the maxima, see figure 6 lower left, contain all the information about the step and

can be seen converging to the location of the step. Indeed, it is possible to reconstruct the input time series from the maxima representation. Let us, however, remove both the maxima lines in question. The reconstructed time series does not possess the step. In fact we see some additional distortion to the time series - at large scales the (almost) constant components in the original function (without the step added) are indistinguishable from the step singularity. This is also reflected in the fact that the maxima do not converge to the locations of the singularities along straight lines. Still the reconstruction (up to the low frequency almost constant bias) illustrates the point: the singularity in the time series is inherited by the maximum line. In fact, as a curiosity, we can remove just one line prior to reconstruction to arrive at 'half a step' reconstructed.

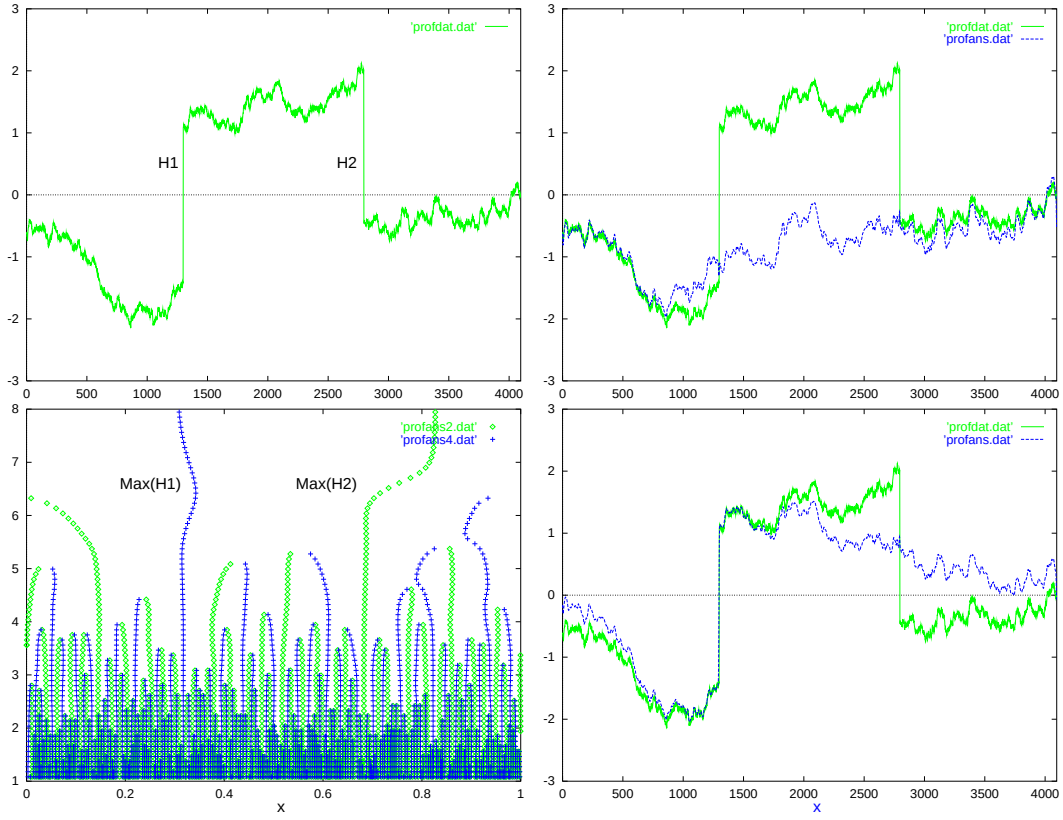

Figure 6: Left above, the input time series; Brownian motion with the step $(H1, H2)$ added. Left below, the WTMM wavelet transform with the first derivative of the Gaussian wavelet. Above right, the reconstruction with two step related maxima lines removed. Below right, the reconstruction with one of the two step related maxima lines removed (Max(H2)).

Alternatively, one could retain only the step related maxima lines and remove the remaining 'noise' in order to reconstruct the step. In the following example, see figure 7, we will retain several maxima lines starting at the largest scales in order to reconstruct the largest singularities in the time series while suppressing the remaining signal, which becomes 'noise' through such a definition.

Even though to each singularity there is a maximum line, provided the number of vanishing moments of the wavelets is sufficient to detect the singularity, at a particular scale only the singularities which are 'large' enough are visible. (Note that this size or strength refers not to the Hölder exponent h, but to the value of the WT maximum at the scale considered.) Therefore, with the scale increasing, we have fewer and fewer singularities visible. We can, therefore, take the set of the largest maxima

lines (those visible at and above some chosen scale) and use them to obtain the approximation of the function to some level of detail (scale), but with the main singularities well focused.

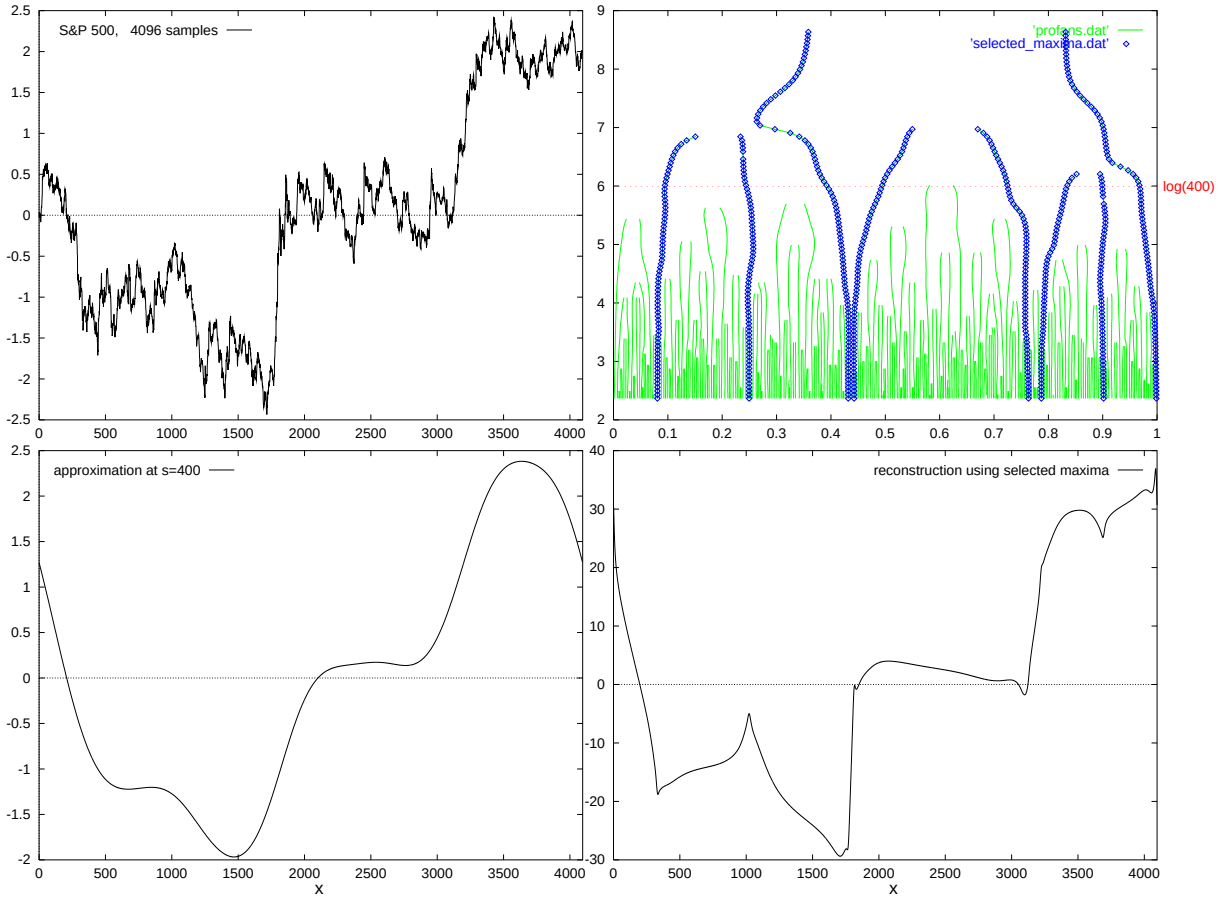

Figure 7: Left above, the input time series: S&P 500 index, first 4096 samples of 16384 shown in figure 1. Right above, the WT maxima obtained with the Mexican hat. Eight selected maxima are highlighted. Lines at scale $s = 400$ and $s = 50$ are drawn. Notice that there are no other maxima except for the highlighted ones, which begin above the scale $\log(400) = 6$. Below left, the low frequency version of the input time series corresponding with smoothing with a Gaussian filter at $s = 400$. Below right, the 'reconstruction' using selected maxima only.

Let us as an illustration refer to figure 7. The input time series is shown together with the tree of maxima of the WT with the Mexican hat wavelet. In the same figure 7 right above, we selected a scale level of $s = 400$ which is denoted with the line. All eight maxima which are still visible at scales larger than the one selected are highlighted and used for 'reconstruction'. In figure 7 below left, we show the low pass approximation of the time series corresponding with the selected scale level. To the right, we show the approximation using the selected maxima. Compared to the smoothing only, the reconstruction gives a much better characterisation of the largest discontinuities, although it shows small overshoots (apparently due to the Gibbs phenomenon).

The detail of the singular behaviour is superimposed on the smooth approximation, as shown in bottom left figure 7. This is evident from the fact that the maxima used are the only ones existing above the selected threshold scale considered, in our case $s = 400$. Note, that the maxima above this

scale fully characterise the approximation at this scale.

This procedure (used for both above examples) is only meant for qualitative illustration purposes; it only works well if one can identify the maxima corresponding to the given singularity. As we have discussed before, the number of lines converging to the singularity depends on the combination of the value of the h exponent and the number of vanishing moments (number of oscillations) of the wavelet. Still, even if the number of maxima is larger than one, each maximum line fully characterises the singularity exponent and its size/scale, even though it may not be sufficient to reconstruct it.

4. Estimation of the Local, Effective Hölder Exponent Using the Multiplicative Cascade Model

We have shown in the previous section 3 that the wavelet transform and in particular its maxima lines can be used in evaluating the Hölder exponent in isolated singularities. The scaling of such singularities remains essentially uniform below some critical scale, making the estimation possible with the linear fit in the log-log plot over a carefully selected scale range.

In most real life situations, however, the singularities in the time series are not isolated but densely packed. The logarithmic rate of increase or decay of the corresponding wavelet transform maximum line is usually not stable but it fluctuates wildly, often making estimation impossible due to divergence problems when the value of the WT along the maximum line approaches zero.

On the other hand, we have also shown on both simulated and real examples that the maxima lines contain the 'compressed' information about the singular behaviour, potentially very relevant for our purpose. Encoding and processing the entire length of the maximum line is relatively computationally expensive, and estimation of Hölder exponent from the log-log fit impossible. Still, we would like to have some means of characterising the singular behaviour from the related maximum line. As a remedy for the estimation problems, we will use the characterisation with the model based approximation of the local scaling exponent, which we will refer to as an *effective* Hölder exponent of the singularity.

In order to estimate this exponent in real life time series with dense singular behaviour, we need to approach the problem of diverging maxima values in log-log plots and the problem of slope fluctuations.

We used the procedure of bounding the local Hölder exponent as described in the report [10] to pre-process the maxima. The crux of the method lies in the explicit calculation of the bounds for the (positive and negative) slope locally in scale. The parts of the maxima lines for which the slope exceeds the bounds imposed are simply not considered in calculations. The output of this procedure is therefore the set of non-diverging values of the maxima lines corresponding to the singularities in the time series.

Even though instead of fluctuating wildly between $+\infty$ and $-\infty$, these values now remain within the bounds, they still fluctuate, with the local slope changing from point to point. Of course, this is why it is not possible to evaluate the Hölder exponent by linear fit in log-log plot, something we can do for isolated cases giving a stable maximum value decay/increase. Therefore, we resort to the second assumption, in which we model the singularities as created in some kind of a collective process of a very generic class. For the estimation of the local Hölder exponent in such time series, we will use a multiplicative cascade model. This will allow us to construct a stable estimate of a local $h(x_0)$ exponent. The multiplicative cascade model is a generalisation of a binomial multiplicative process otherwise known as the Besicovich binomial process.

4.1 Multiplicative Cascade Model

The Besicovitch measure is actually a simple extension of the widely known Cantor set construction achieved by equipping it with a multiplicative measure. To demonstrate a uniform case, we start from a unit mass bar uniformly distributed over (0..1) interval. In the first generation step, the support is divided into three equal parts and the unit mass is divided in two and distributed over the side intervals of length 1/3, being (0..1/3) and (2/3..1). Note that the centre interval remains empty. In the next generation, the same procedure is recursively applied to all the intervals with mass distributed over them. It is easy to check that each step of generation increases the density of the measure by the

factor 3/2, while the total measure remains constant and equal to the unit mass.

There is the possibility of generalising this construction through non-equal factors defining non-uniform, multiplicative repartitioning of the measure. To do this, one again takes a unit measure and distributes it with the arbitrary ratios p_1 and p_2 over the two remaining sections of the line at each construction step, see figure 8 left. The resulting *multiplicative* distribution of the measure gives a classical example of the so-called *multi-fractal* object [11].

Naturally, the ratios $c_1^{-1} = c_2^{-1} = 1/3$ defining the middle-third Cantor set can as well be set to non-uniform. Also, the number of divisions, which is equivalent to the number of transformations, see Eq. 4.1, can be subject to alteration (increase). In particular the support does not have to be the Cantor set at all, but it can simply be the entire (0..1) interval. If in addition to this the normalisation requirement is lifted, we will refer to such construction as the multiplicative cascade model, see figure 8 right.

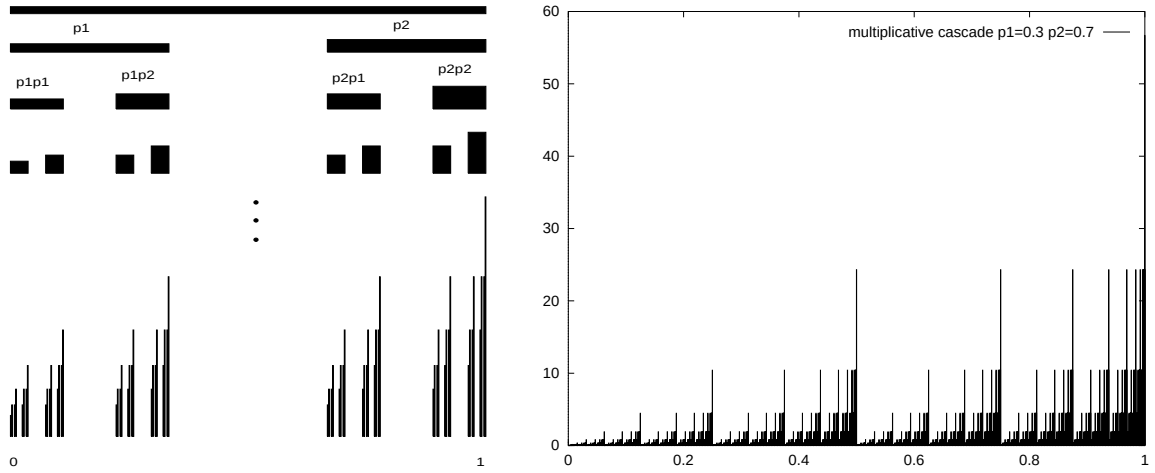

Figure 8: Left: the Besicovitch measure on the Cantor set, generations F_0 through F_3 and the generation F_6. The distribution of weights is $p_1 = 0.4$ and $p_2 = 0.6$. The standard middle third Cantor division is retained. Right: similar construction but on 0..1 support instead of the Cantor set, leading to multiplicative cascade. $p_1 = 0.3$ and $p_2 = 0.7$, generation F_{13}.

The set of transformations $B_{\{1,2\}}$ describing the Besicovitch construction can be expressed as:

$$\begin{aligned} B_1\ f(x) &= p_1\ f(\frac{x+b_1}{c_1})\ ; \\ B_2\ f(x) &= p_2\ f(\frac{x+b_2}{c_2})\ , \end{aligned} \tag{4.1}$$

with the normalisation requirement:

$$p_1 + p_2 = 1\ . \tag{4.2}$$

Additionally, we put conditions ensuring non-overlapping of the transformations:

$$\frac{1+b_1}{c_1} < \frac{0+b_2}{c_2}$$

while all the respective values $b_1/c_1, b_2/c_2, c_1^{-1}, c_2^{-1}$ are from the interval $(0, 1)$.

For equal ratios, $p_1 = p_2 = 1/2$ and $c_1 = c_2 = 3$ with $b_1 = 0$ and $b_2 = 2$ we recover the middle-third, homogeneous distribution of measure on the Cantor set. We have the Besicovitch measure for non-equal p_i, with other above settings retained. Finally, for non-equal p_i, regardless of normalisation Eq 4.2 and with $c_1 = c_2 = 2$ with $b_1 = 0$ and $b_2 = 1$, we have the multiplicative cascade on (0..1) interval.

Each point of this cascade is uniquely characterised by the sequence of weights $(s_1...s_n)$ taking values from the (binary) set $\{1, 2\}$, and acting successively along a unique process branch leading to this point. Suppose that we denote the density of the cascade at the generation level F_i by $\kappa(F_i)$, we then have

$$\kappa(F_{max}) = p_{s_1} \; ... \; p_{s_n} \; \kappa(F_0) = P_{F_0}^{F_{max}} \; \kappa(F_0)$$

and the local exponent is related to the product $P_{F_0}^{F_{max}}$ of these weights:

$$h_{F_{max}}^{F_0} = \frac{\log(P_{F_0}^{F_{max}})}{\log((1/2)^{max}) - \log((1/2)^0)} \; .$$

In any experimental situation, the weights p_i are not known and h_i has to be estimated. This can be simply done using the fact that for the multiplicative cascade process of the kind just described, the effective product of the weighting factors is reflected in the difference of logarithmic values of the densities at F_0 and F_{max} along the process branch:

$$h_{F_{max}}^{F_0} = \frac{\log(\kappa(F_{max})) - \log(\kappa(s_0))}{\log((1/2)^{max}) - \log((1/2)^0)} \; .$$

The densities along the process branch can be estimated with the wavelet transform using its remarkable ability to reveal the entire process tree of a multiplicative process [9, 12]. It can be shown that the densities $\kappa(F_i)$ can be estimated from the value of the wavelet transform along the maxima lines corresponding to the given process branch. The estimate of the effective Hölder exponent becomes:

$$\hat{h}_{s_{min}}^{s_{max}} = \frac{\log(Wf\omega_{pb}(s_{min})) - \log(Wf\omega_{pb}(s_{max}))}{\log(s_{min}) - \log(s_{max})}$$

where $Wf\omega_{pb}(s)$ is the value of the wavelet transform at the scale s, along the maximum line ω_{pb} corresponding to the given process branch. Scale s_{min} corresponds with generation F_{max}, while s_{max} corresponds with generation F_0.

For the estimation of h, we need s_{max} and $Wf\omega_{pb}(s_{max})$. We can, of course, pick any of the roots of the sub-trees of the entire maxima tree in order to evaluate exponents of the partial process or sub-cascade. But for the entire sample available we must use the entire tree and for this purpose, we can only do as well as taking the sample length to correspond with s_{max}, i.e.:

$$s_{max} \equiv s_{SL} = \log(SampleLength) \; .$$

Unfortunately, the wavelet transform coefficients at this scale are heavily distorted by finite size effects. This is why we estimate the value of $Wf\omega_{pb}(s_{max})$ using the mean h exponent.

4.2 Estimation of the Mean Hölder Exponent

For a multiplicative cascade process, a mean value of the cascade at the scale s can be defined as:

$$\mathcal{M}(s) = \frac{\mathcal{Z}(s,1)}{\mathcal{Z}(s,0)} \,, \tag{4.3}$$

where the $\mathcal{Z}(s,q)$ is the partition function of the q-th moment of the measure distributed over the wavelet transform maxima at the scale s considered:

$$\mathcal{Z}(s,q) = \sum_{\Omega(s)} (Wf\omega_i(s))^q \,,$$

where $\Omega(s) = \{\omega_i(s)\}$ is the set of all maxima $\omega_i(s)$ at the scale s, satisfying the constraint on their local logarithmic derivative in scale [10]. This mean gives the direct possibility of estimating the mean value of the local Hölder exponent as a linear fit to M:

$$\log(\mathcal{M}(s)) = \bar{h} \log s + C \,. \tag{4.4}$$

We will not, however, use the definition 4.3 since we want the Hölder exponent to be the local version of the Hurst exponent. This compatibility is easily achieved when we take the second moment of the partition function to define the mean $\bar{h}'$:

$$\mathcal{M}'(s) = \sqrt{\frac{\mathcal{Z}(s,2)}{\mathcal{Z}(s,0)}} \,.$$

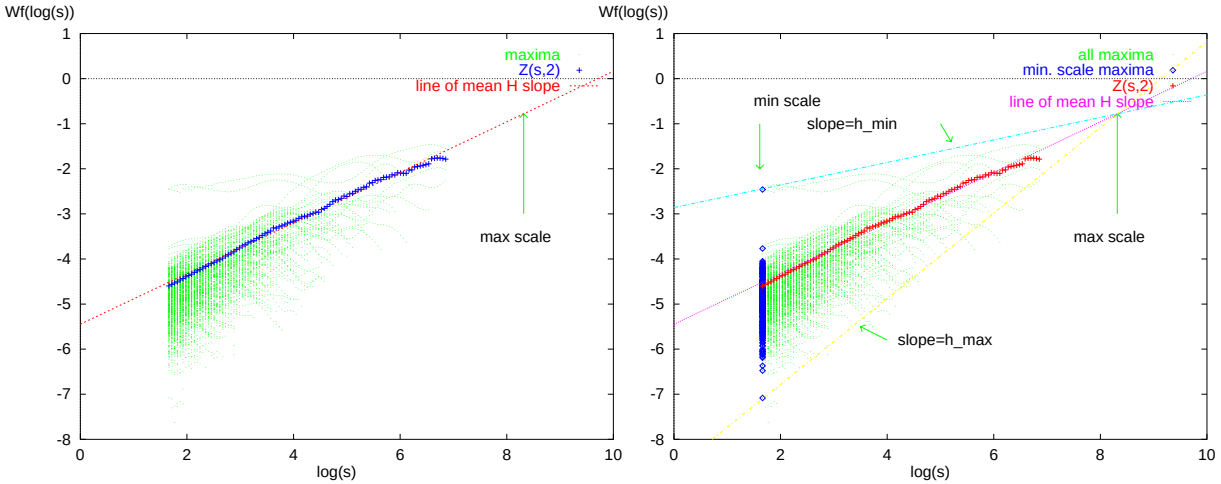

Figure 9: Left: the projection of the maxima lines of the WT along time. The mean value of the Hölder exponent can be estimated from the log-log slope of the line shown. Also, the beginning of the cascade at the maximum scale s_{max} is indicated. Right: the maxima at the smallest scale considered are shown in the projection along time. The effective Hölder exponent can be evaluated for each point of the maximum line at s_{min} scale. Two extremal exponent values are indicated, for minimum and maximum slope.

Therefore, we estimate our mean Hölder exponent $\bar{h}'$ from 4.4 substituting $\mathcal{M}$ with $\mathcal{M}'$. The estimate of the local Hölder exponent now becomes:

$$\hat{h}^{s_{SL}}_{s_{min}} \cong \frac{\log(Wf(s_{min})) - (\bar{h}' \log s + C)}{\log(s_{min}) - \log(s_{SL})} \ .$$

4.3 Comparing the Distributions and the Evolution of the Logarithmic Histogram
Such an estimated local $h(x_0, s)$ can be depicted in the temporal fashion, for example with colour/grayscale as we have done in figure 1. Alternatively it can be grouped into histograms. We will estimate histograms of $h(x_0, s)$ for a *range* of scales less than factor 2.0; $s_{min}/s_{max} < 2.0$. This is done in order to increase the number of points to be histogrammed - instead of one scale, we sample a multitude of scales within a narrow scale range (10 samples per scale for histograms shown in figure 10).

We display histograms of h, taking the logarithm of the measure in each histogram bin. This conserves the monotonicity of the original histogram, but allows us to compare the log-histograms with the so-called *spectrum of singularities* $D(h)$.[3] It is a standard way of visualising the distribution of singularities. It gives the (fractal) dimension $D(h_i)$ of the supporting set of singularities for each exponent value h_i in the time-series. This is usually obtained using the so-called Legendre Transform from the moments of the partition function $\mathcal{Z}$, see Ref[13, 14], but there is also a direct correspondence between our log-histograms and the $D(h)$ through the scaling of the logarithmic histograms:

$$D(h) = dim(\{x_0\} \ : \ T(x - x_0) \sim |x - x_0|^{h(x_0)}) \sim \frac{\log(\mu(h(s_{max}))) - \log(\mu(h(s_{min})))}{\log(s_{max}) - \log(s_{min})} \ .$$

For three example time series, we show in figure 10, log-histograms of the exponent h at different scales. The time series considered are a white noise sample, a fractional Brownian motion with $H = 0.6$, and a record of the S&P index.

Starting at the top, the row of histograms is made for the scale range $\log(500.0) < \log(s) < log(1000.0)$. The histograms show considerable fragmentation. Several modes become visible and in an extreme interpretation, all the values can be considered as single modes. This will certainly hold for even higher scales with the limit of one single value. This limit is achieved in less than a decade of scale for the time series shown - the 'histogram' plots (not shown) for the scale $\log(s) = 6.53$ have only one value. This means that for $\log(s) = 5.15$ and all the scales above, the fluctuations dominate the distribution and consistent statistical behaviour becomes dispersed. On the contrary, while going down with the scale, the bulk of consistent behaviour overshadows the large scale fluctuations. This can be observed in the second row of histograms for the scale $\log(50.0) < \log(s) < log(100.0)$, and especially for the fourth row at the scale $\log(5.0) < \log(s) < \log(10.0)$.

The consistent statistical behaviour is captured in the scale-free representation of these histograms. The $D(h)$ spectrum provides such a representation, capturing both gross, scale-free behaviour and the fluctuations (giving rise to the off-centre spreading of the spectrum). In the ideal case of infinite sample length, it should be just one Dirac delta at some value h for the first two time series from the left. Evidently this is not the case and the fluctuations of the relatively short length time series sample (4096 points) are the reason for this. While this is, of course, the reason for taking longer records for deriving statistically meaningful estimates, we will show in the following that, for the task of comparing time series, e.g. various realisations of random processes, with respect to similarity criteria, the fluctuations constitute the core elements for such a comparison.

With the h-exponent scale parameterized histograms, we have therefore obtained insight into both the largest events in scale (fluctuations) and largest events in h distribution (the rare events). Both can be used as a means of characterising time series, along with the distribution at a certain scale or the invariant limit distribution $D(h)$. Since we aim at a local comparison, we will not expand on the issue of the distribution matching, and will move directly to the issue of local representation.

[3] this is usually denoted with $f(\alpha)$ in the literature, but we find $D(h)$ more suitable here.

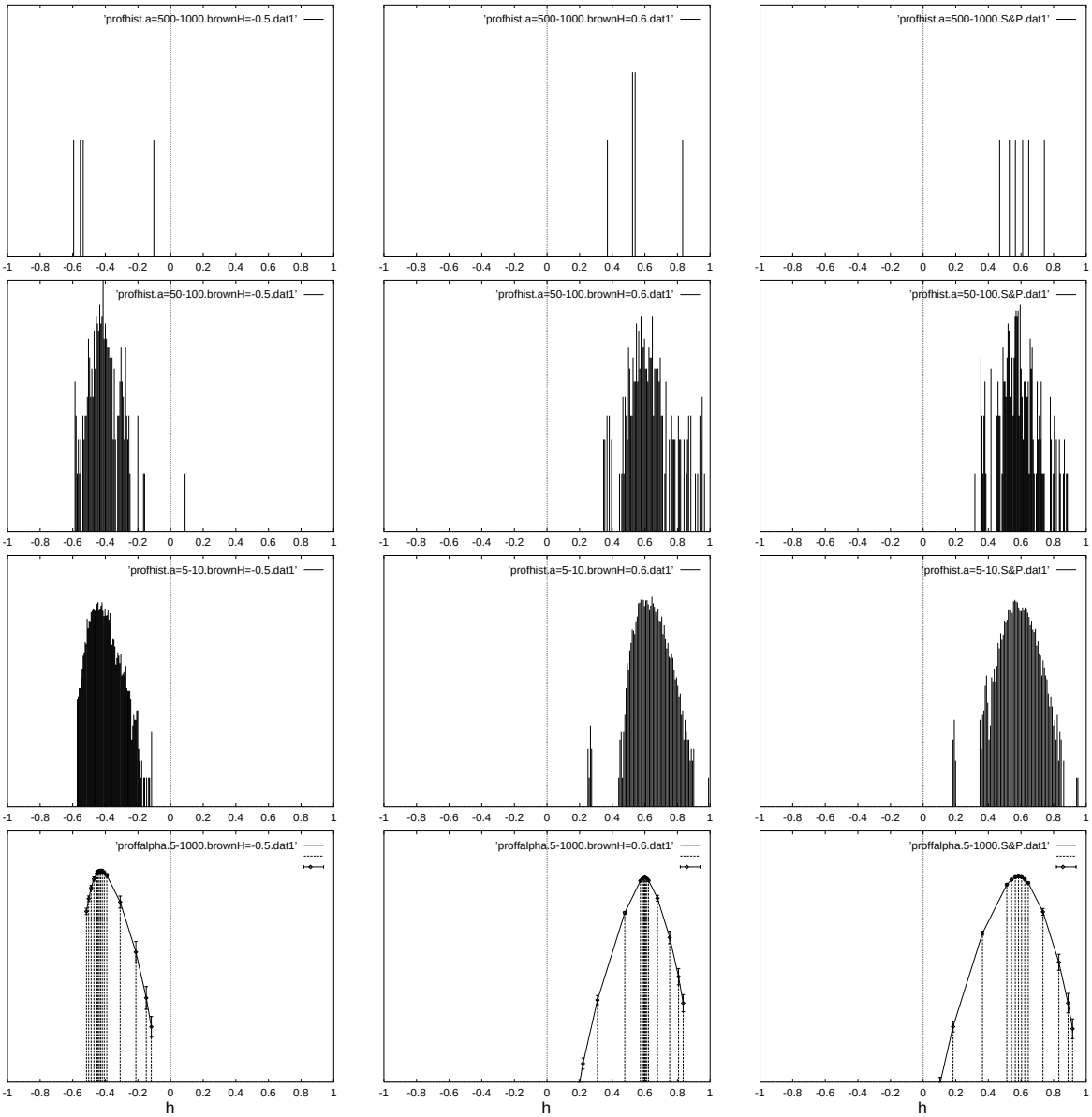

Figure 10: Three sets of h histograms for respective scales $\log(s) = 5.15$, $\log(s) = 3.75$, $\log(s) = 2.31$, respectively for the top first, second and third row. Below, in the bottom row, the corresponding $D(h)$ spectrum. Left column: for 4096 samples of white noise. Centre: 4096 samples of fractional Brownian motion with $H = 0.6$. Right: S&P 500 index, first 4096 samples from figure 1.

5. The h-Representation

As already discussed in section 3, the wavelet transform removes the polynomial bias, but at the same time it effectively 'compresses' the information about the 'non-stationarity' into a piece of local information. Moreover, it reveals the scale-wise organisation of singularities, thus allowing for the selection of the interesting strongest events. This we have discussed in the previous section; the strongest events will slowly disappear in the bulk of the maxima while going down the scale.

In oder to arrive at a (very compact) representation of the time series, one would like to include a certain (predefined) number of such features in it. Therefore, one would have to find an optimum of the scale of representation and the number of features, a process prone to some arbitrariness in the design of optimality criteria. This problem can be avoided using a somewhat modified strategy which we suggest in the following. The h-representation as we will call it will be obtained by means of tracing the fixed number of strongest maxima *below* the representation scale at which they appear, thus allowing better localisation of singular features in the time domain and a more stable estimation of the h exponent.

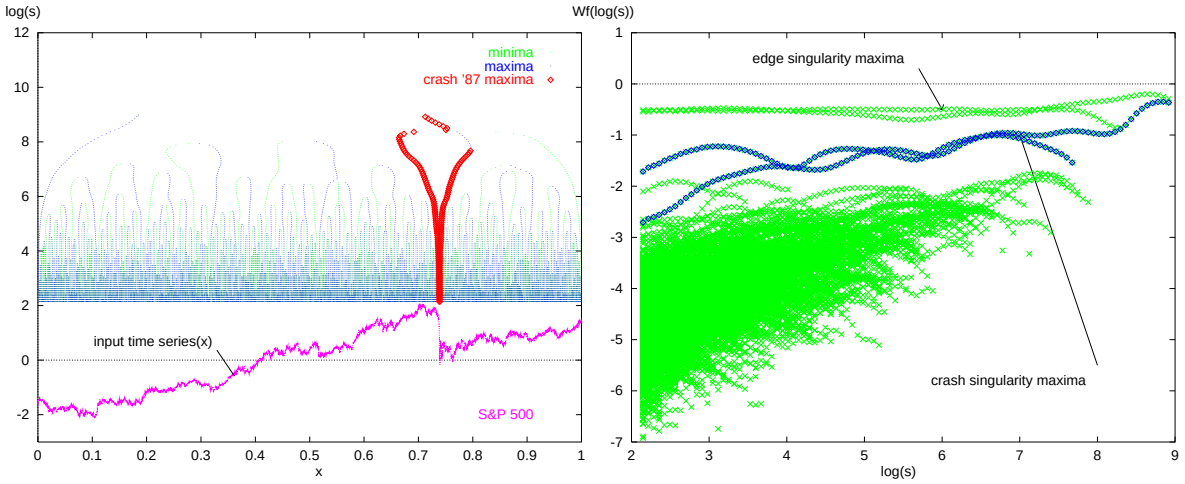

Figure 11: Left: the input time series with the WT maxima above in the same figure. The strongest maxima correspond to the crash of '87. The input time series is de-biased and L1 normalised. Right: we show the same crash related maxima highlighted in the projection showing the logarithmic scaling of all the maxima.

Let us consider the entire sample length of the S&P 500 index shown in the lower part of figure 1. We can make the decomposition with the Mexican hat wavelet. We will use the Mexican hat, since it will mark with one line the places where we have 'a change of slope'. Of course, the step Heaviside function will have two lines approaching it, and the Dirac delta three.

The strongest maxima in figure 11 left (above the input time series plot) converge to the largest singular events. Note that the largest singularity is *not* related to the highest amplitude of the time series but to the largest step like singularity. Also, another set of strongest maxima corresponds to the step singularities at the end points of the time series resulting from the finite length of the investigated time series.

In the right figure 11, we show the same maxima highlighted in the projection showing the logarithmic scaling of all the maxima. The maxima corresponding to the crash are the strongest for all the scales considered. They also show scaling which is closer to that of the Heaviside step with Hölder exponent $h = 0$, rather than to the average Hurst exponent of the financial time series $H \sim 0.5$.

Let us now show the development of h values associated with each maximum for some meaningful range of scales, e.g. from $\log(s) = 2$ to $\log(s) = 8$. Even in plotting the projection of the entire set of h values, several lines show up as the largest deviations, see figure 12 left. They extend far from the

main 'cone' of the distribution and can be identified as: a) the maxima corresponding with the ends of the time series record and b) the maxima related to the strongest event, the crash of '87. Taking a section of the projection of h for a fixed scale confirms this observation. In figure 12 right, a section is carried out for $\log(s) = 6$ above, and for $\log(s) = 2.14$ below. For each value of the h exponent, an impulse is drawn; what we show is therefore not a histogram but it contains the same information; the histogram can be obtained from it by appropriately binning the data.

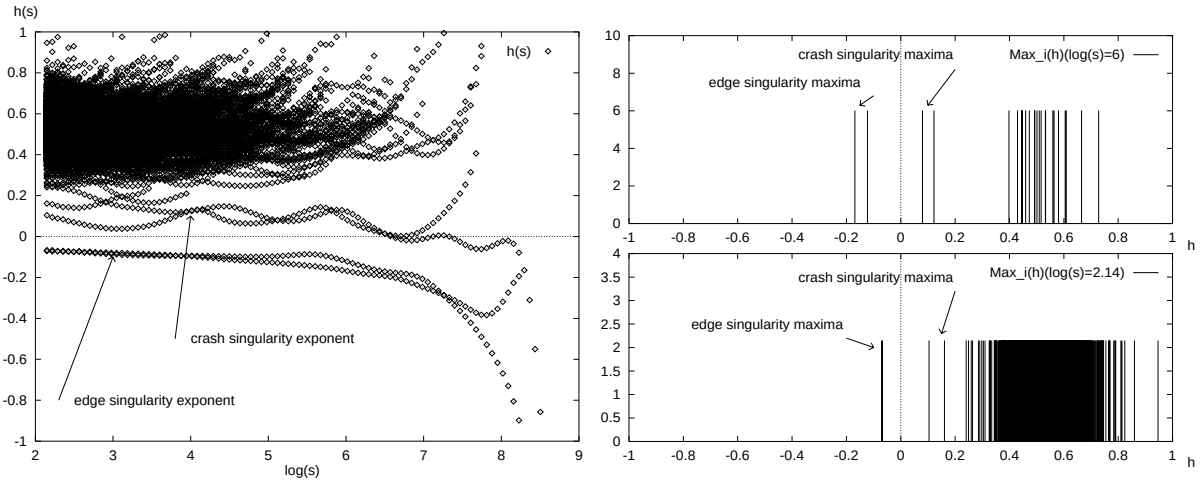

Figure 12: Left: the projection of the local Hölder exponents along scale. The exponents of the major events, edge and crash are indicated. Right: the distributions of the local Hölder exponents taken at two different scales, $\log(s) = 6.0$ and $\log(s) = 2.14$.

We chose this way of presentation since the strongest events on which we want to focus are not very frequent - we refer to them as 'rare' events. They are difficult to see once the bulk of the distribution takes shape, as in the section at $\log(s) = 2.14$ in figure 12 bottom right.

Note that at large scales, the crash related maxima are evidently and strongly unique, while at smaller scales they loose the distance to the bulk of maxima lines. This observation suggests that one should look for some optimal combination of scale and number of the strongest features to be represented. Therefore, the main criterium for selecting the working scale used would be the predefined small number of features which are revealed up to this scale. This approach is, however, not without problems - the parameters, like position x_i and the Hölder exponent h_i, obtained for such a relatively high working scale would be burdened with a very big distortion. This is why we use a somewhat modified strategy. We evaluate the maxima decomposition for some considerable range of scales, say two decades or more, and select the predefined number of maxima which show up first, while going down the scale. The relevant parameters can thus be evaluated for these maxima over the extended range of scales.

Let us illustrate this by selecting the 10 strongest maxima from the distribution in figure 12 and continuing them down to scale $\log(s) = 6$. They are indicated by points in figure 13 left, while in top right, the corresponding 'distribution' of h_i values is shown. Finally, in the same figure 13, we show the temporal structure of the analysed time series sampled along the chosen set of points; it is a set $\{x_i, f(x_i)\}$, where i runs from $1, ..., 10$, enumerating the strongest maxima chosen. For the sake of comparison, we plot in figure 14 the sampling of the input time series with all the maxima (which are 25) present at the scale $\log(s) = 6$. There is a substantial amount of detail added to the 'approximation', nevertheless the strongest features remain unchanged. In figure 13 right, we compare the sampling with the 10 strongest maxima against the original time series. We determine that the conclusion is not different - the largest features are well captured by the sampling proposed.

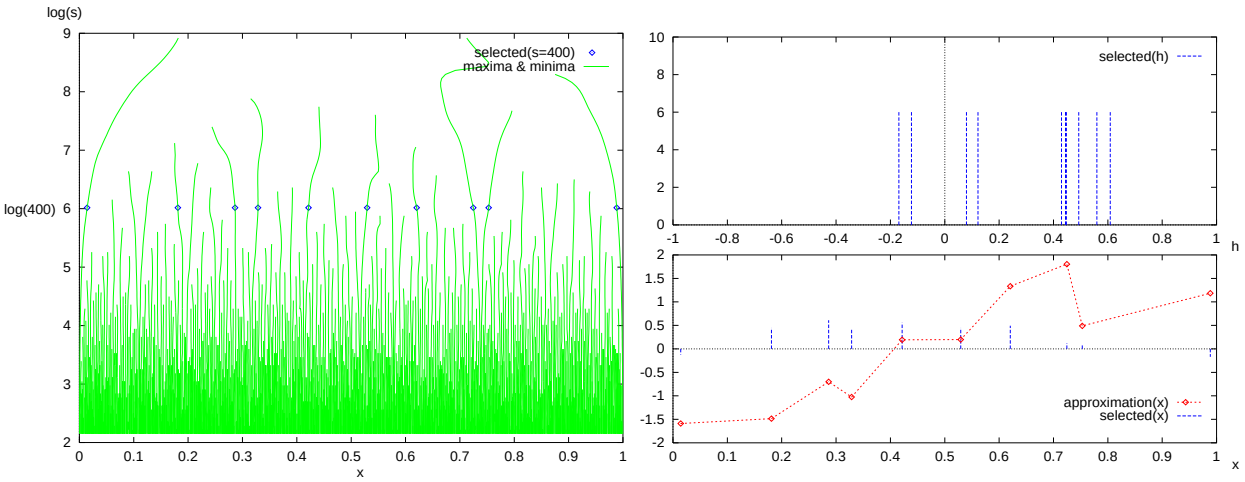

Figure 13: Left: the points indicating the strongest 10 maxima at the lowest scale considered. Right above: the distribution of the h values. It retains the rare events, and shows far fewer bulk events. Right below: the sequence of h in the temporal fashion. Also in the same plot, we show the 'approximation' of the time series using the sampling of the original time series at the locations of the selected maxima.

Note that it is not the values of the function which are retained for the sake of representing the time series, but the corresponding (effective) Hölder exponent. Indeed, generally we would not want to be dependent on the exact values of the time series, but rather employ the scale free characteristics, locally independent of vertical rescaling and polynomial bias. Even though we discard the actual values of the wavelet transform at the chosen maxima points, the signs of these values can be taken into the representation. Generally speaking, they will allow us to distinguish between inverted features and, in particular, between the time series $f(x)$ and its inverted version $-f(x)$.

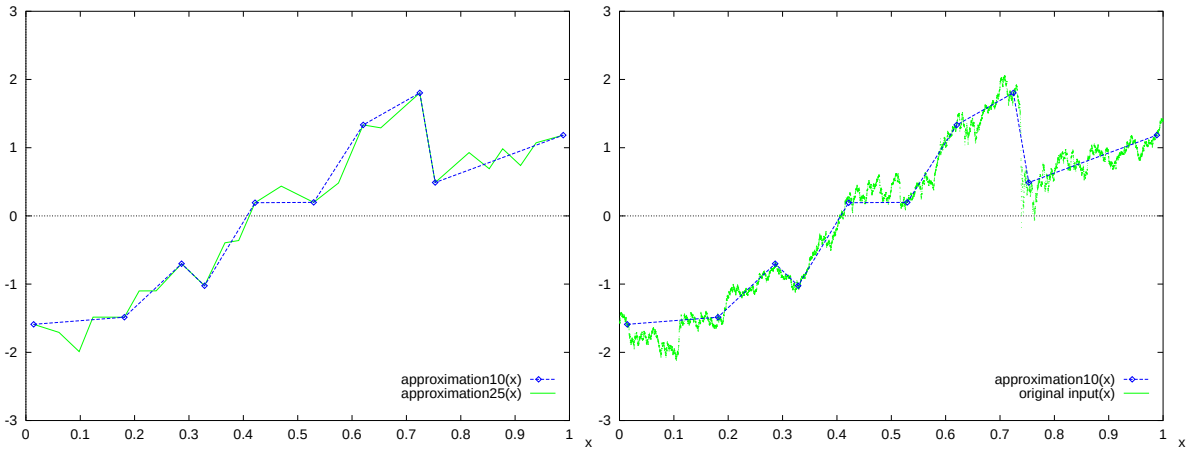

Figure 14: Left: the 'approximation' of the time series using the 10 strongest maxima, overlayed onto the 'approximation' using all 25 maxima at the scale considered $\log(s) = 6$. Right: the 'approximation' of the time series using the strongest 10 maxima, overlayed onto the original time series.

In figure 15 left, we show the sequence of the values of the 10 strongest maxima at the scale considered. The signs of these values give a unique sequence of sign changes. This information is not present in the sequence of the local Hölder exponent as shown in figure 13 bottom right. This is due to the fact that we take the logarithm of the modulus of the WT value to estimate the scaling exponent. Still, this sign sequence provides important information about the sign of the singularity in addition to its exponent. There is a unique relation between the sign of the WT and the curvature/concavity of the approximation pattern - compare the left figures 15 and 14. The reason for this is evident from the fact that this sign sequence simply reflects the sign of the second derivative of the smoothed version of the time series (for the Mexican hat wavelet). The distribution of these maxima is almost symmetrical around zero, which can be seen in figure 15 right.

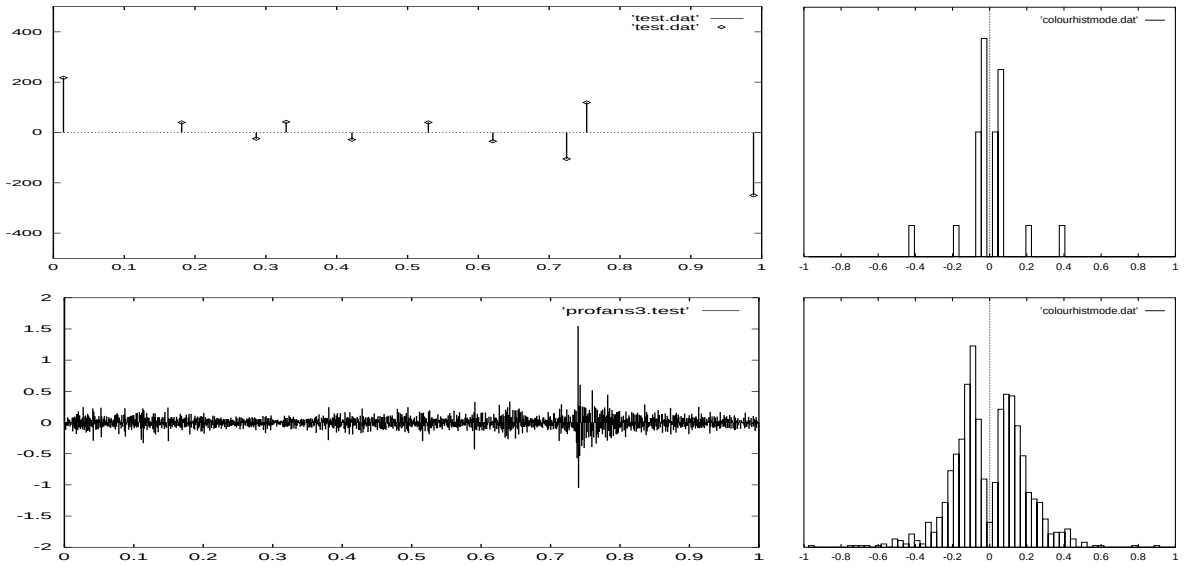

Figure 15: Left above: the sequence of the values of the 10 strongest maxima at the scale $\log(s) = 6$, giving a unique sequence of sign changes. Right above: the distribution of these maxima values is almost symmetrical around zero. Below: analogical plots for scale $\log(s) = 2.14$.

In conclusion to the above considerations, we can design our h-representation to contain the set of a certain number of the largest features of the time series at hand. The parameters coded are the x-coordinate x_i of the selected maximum lines ω_i at the scale s_{min}, the Hölder exponent $h(x_i)$, the corresponding sign of the wavelet transform $Wf(x_i)$ (and optionally the top scale of ω_i). The implicit assumption is taken here, which allows us to neglect the top scale parameter within the default h-representation. It relies on the expectation that for the small number of largest features their top scale will not differ much, and most likely will be within less than one decade. It is of course possible to validate this assumption in the process of constructing the representation, as is also possible to simply retain the top scale and take it into the representation. Within the scope of this work, however, we do not take the top-scale parameter into consideration when deriving the h-representation and as a consequence when comparing the time series in the test section 6.

5.1 Matching Distances and Norms

We used a straightforward algorithm to evaluate the similarity between the h-representations. For each set of numbers associated with the representation feature i, we used a quadratic distance measure with separate factors for position and h exponent, f_x and f_h respectively:

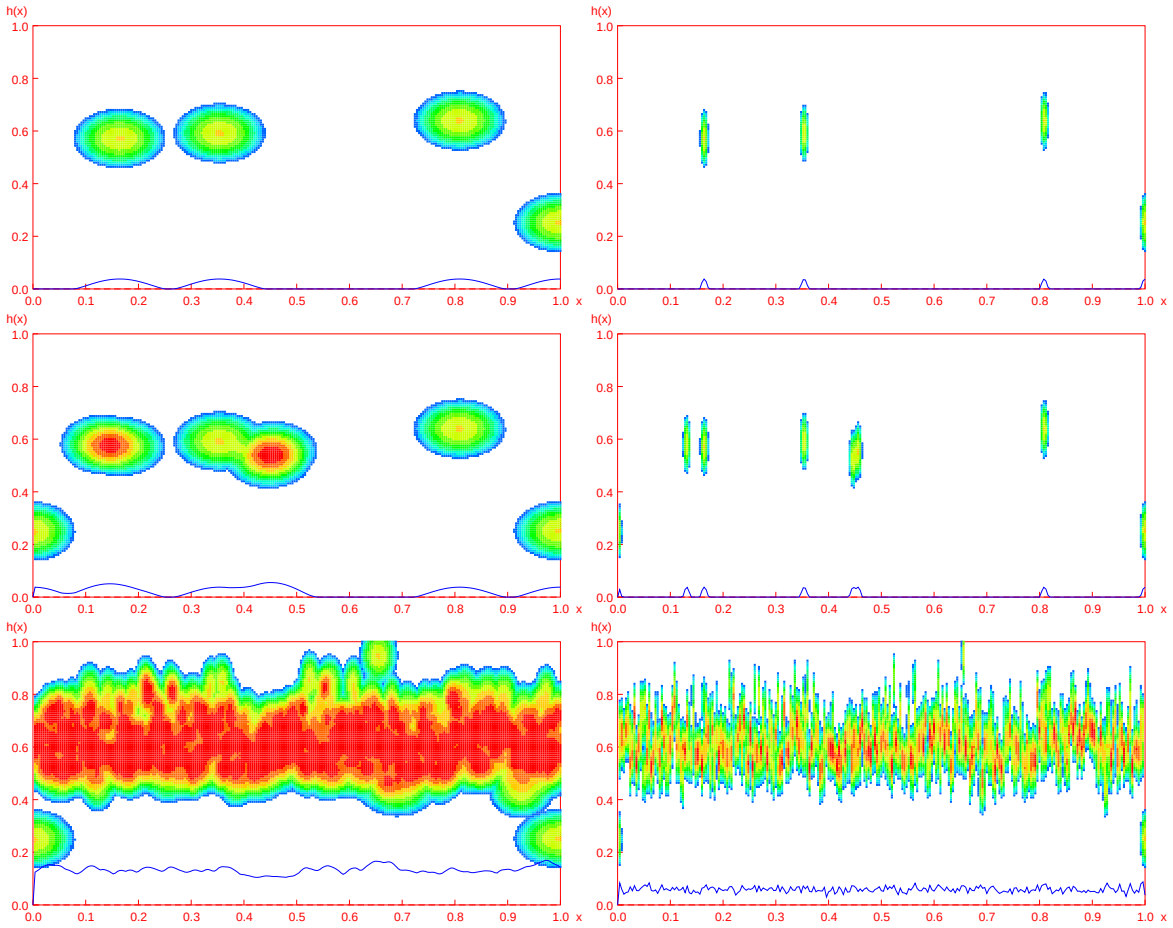

Figure 16: Two sets of autocorrelation measures $corr_{s,s}(x,h)$ and $Corr_{s,s}(x)$ for different f_x factors and different numbers of features. From top to bottom 4, 8 and 547 features, and $f_x = 1.0$ for the left column and $f_x = 0.1$ for the right column of the plots. The pointwise autocorrelation $corr_{s,s}(x,h)$ reflects distance measure, and is shown with colour changing from red to blue according to distance increasing. The autocorrelation product $Corr_{s,s}(x)$ is shown as a line plot function of time x. One can verify that it corresponds with the vertical projection of the distance measure.

$$dist_s(x,h) = 1 - (f_x \, {\Delta_x}^2 + f_h \, {\Delta_h}^2) \, ,$$

where $\Delta_x = x - x_i$ and $\Delta_h = h - h_i$ and x_i, h_i belong to s - the representation of the time series.

The representation thus defined is suitable for determining the distance measure between the time series. A simple pointwise product will show how the two representations s_1 and s_2, of the time series in hand are correlated in the time x, and h exponent domains:

$$corr_{s_1,s_2}(x,h) = dist_{s_1}(x,h)\, dist_{s_2}(x,h) \, . \tag{5.1}$$

Since we actually have the h-representation consisting of two (independent) parts, corresponding with the positive and the negative signs, we also have two distance functions per time series. This is why we use the 'modulus' of the two part correlation function:

$$|corr^{+-}_{s_1,s_2}(x,h)| = \sqrt{\frac{1}{2}}\,\sqrt{dist^{+}_{s_1}(x,h)\;dist^{+}_{s_2}(x,h) + dist^{-}_{s_1}(x,h)\;dist^{-}_{s_2}(x,h)}\;. \tag{5.2}$$

Calculation of the actual correlation product in time domain can now be done by a simple projection of the pointwise correlation onto the time axis x:

$$Corr_{s_1,s_2}(x) = \int_{h_{min}}^{h_{max}} |corr^{+-}_{s_1,s_2}(x,h)|\;dh\;. \tag{5.3}$$

From this time dependent correlation product we can, of course, evaluate the cumulative correlation measure for comparison between different pairs of time series, i.e. for determining the similarity measure. For this purpose, however, we have to normalise the correlation measure. The simplest norm can be evaluated just as the autocorrelation of the h-representation of the time series s:

$$N_s = \int_0^1 Corr_{s,s}(x)\;dx\;, \tag{5.4}$$

where we assume that the time span x has been normalised to the 0..1 range. Therefore the measure of similarity between time series s_1 and s_2 can be evaluated as:

$$sim_{s_1,s_2} = \frac{1}{N_{s_1}\;N_{s_2}}\;\int_0^1 Corr_{s_1,s_2}(x)\;dx\;.$$

This similarity measure takes values from the [0..1] interval. Of course, for a distance measure, we can use $-\log(sim)$, extending from 0 to ∞.

5.2 Transition from Local to Stochastic Representation and the Saturation of h-*Representation*
In figure 16 we show an example of pointwise autocorrelation $|corr^{+-}_{s,s}(x,h)|$ of the h-representation of a time series s for two values of f_x factor and for three levels of the number of features represented. We can see how (for both factor f_x values), while incorporating more and more features, the distance measure fills the entire interval (0..1). This happens due to increasingly dense coverage of the position coordinate with more and more features. Thus, the position discriminating capability becomes lost, while the h-discriminating ability approaches that of the $D(h)$ distribution discussed earlier.

The number of features taken into the h-representation can be directly related to the scale; the higher the number of the features the lower the scale s. Therefore, the effect just described is analogical to the histogram 'saturation' which we observed in section 4. Here, it is the saturation of the local h-representation, if we change the scale towards a higher resolution. This can be measured using the total norm measure Eq 5.4, i.e. the cumulative autocorrelation. For the plot of the norm as a function of scale, for two f_x factor values, see figure 17. The evolution of the total norm is linear with respect to $\log(s)$ for the larger sigma factor, and changes to linear from apparently exponential (or power law) for the smaller sigma factor. The local h-representation can in principle be used across the entire spectrum of the scale range and position parameters. The total autocorrelation indeed grows and this reflects the increasing discrimination power with the larger scale range covered. But linear in $\log(s)$, the growth of discrimination power is much slower than the exponential number of features added, and soon it becomes infeasible to use it. This is another indication of the fact that the representation so defined only makes sense for a very small number of features. Indeed, this is perfectly consistent with the design criteria we imposed. Nevertheless, it is possible to extend the h-representation over scale information (top-scale) for those applications requiring larger number of features and scale-wise resolution power.

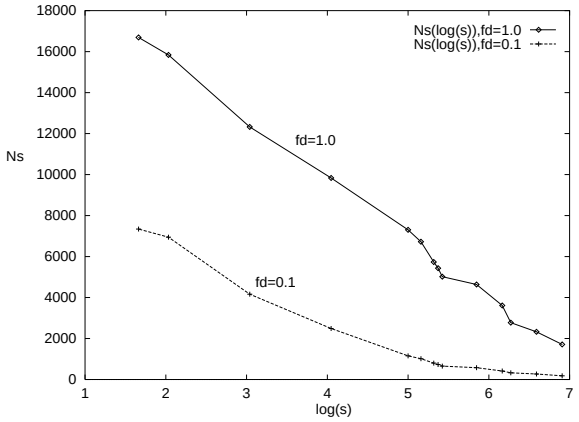

Figure 17: The evolution of the norm N_s as a function of the logarithm of scale, for two different factors $f_x = 1.0$ and $f_x = 0.1$.

6. Experiments with Similarity

We took the records of the exchange rate with respect to USD over the period 01/06/73 - 21/05/87. It contains daily records of the exchange rates of five currencies with respect to USD: Pound Sterling, Canadian Dollar, German Mark, Japanese Yen and Swiss Franc. (Some records were missing - we used the last known value to interpolate missing values.) Below, in figure 18 we show the plots of the records.

All the time series were decomposed using the Mexican hat wavelet. For each, the $10-20$ strongest maxima were selected and for each of these maxima, the following were retained: the position of the maximum at the fine scale, the estimate of the Hölder exponent, the sign of the WT value at the location of the maximum at the finest scale.

As the measure of similarity for our examples, we have respectively:

- German Mark(s_3) versus Swiss Franc(s_5); total correlation = 0.793370
- Pound Sterling(s_1) versus Canadian Dollar(s_2); total correlation = 0.287755
- Pound Sterling(s_1) versus German Mark(s_3); total correlation = 0.408833
- Pound Sterling(s_1) versus Swiss Franc(s_5); total correlation = 0.375356
- Canadian Dollar(s_2) versus German Mark(s_3); total correlation = 0.314108
- Canadian Dollar(s_2) versus Swiss Franc(s_5); total correlation = 0.337519 .

Note that all these values were obtained including the end cut-off and the related singularity at the beginning and at the end of the time series record. (We had to pad with zeros in order to obtain power of 2 for FFT). These cut-off singularities are trivially correlated for all time series and add some bias to the correlation values. For all the above examples, the cut-off singularities account for some $0.1 - 0.2$ correlation.

We plot the time series pairwise, their corresponding h-representations and their pairwise correlation in figures 19 to 21. For the plots of h-representations, we actually plotted the pointwise autocorrelations, which reflects the distance measure. The pointwise correlations of the corresponding h-representations are shown in the bottom plots, including the projection of the pointwise correlation onto the time axis, shown with the line plot. Even at the very low resolution of the h-representations, the correlation plot conveys relevant temporal information about the local similarity of time series

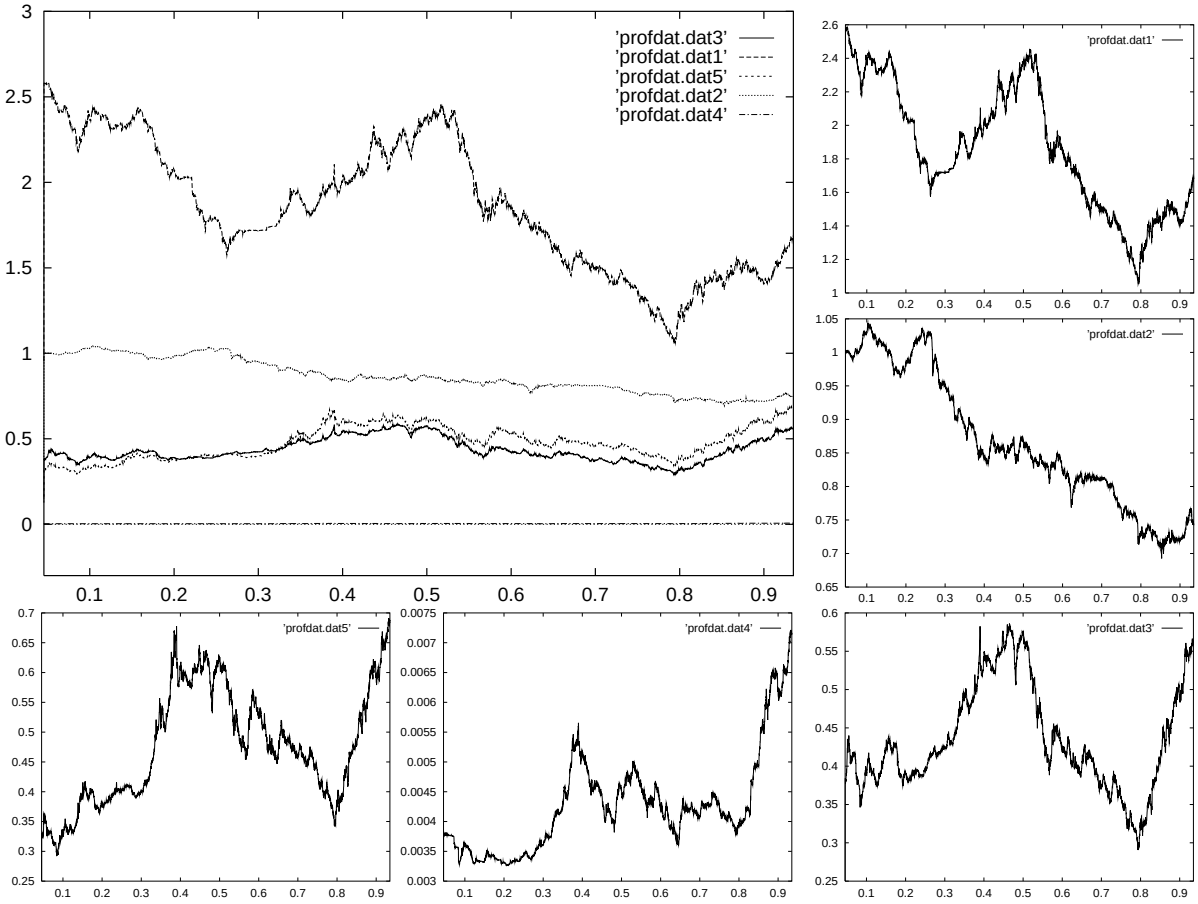

Figure 18: Left above, all the records of the exchange rate used, with respect to USD over the period 01/06/73 - 21/05/87. In small inserts, single exchange rates renormalised, from top right to bottom left (clockwise), Pound Sterling, Canadian Dollar, German Mark, Japanese Yen and Swiss Franc, all with respect to USD.

matched. For example, in figure 19 the time series s_3 and s_5 correlate very well across the entire sample. The time series s_1 and s_2 only start to show some significant correlation after $x = 0.55$ on the normalised time axis. Similarly, in figure 20 we identify a lack of local correlation between $x = 0.2$ and $x = 0.55$.

A possible interpretation is that the time series s_3 and s_5 are permanently strongly coupled through some political/economical links. Considering these are both time series from the European Union, this is not an unlikely reason. On the other hand, the localised beginning of the correlations between the s_1 and s_2 time series may have something to do with an important political/economical/military event which then took place and has coupled both currency systems since then. Alternatively, and perhaps even more likely, the events reflected by both the exchange rates of the currencies in question may have primarily affected the reference currency, in this case the USD.

As an additional test, we measured the similarity between one and the same time series, but with the time reversed in one of them. The results were in all cases convincing - a very low global correlation level. Locally the residual correlation measure was symmetrical. This confirms the fact that there is little long-range correlation (if any) in the records of exchange rates. Still, we can encounter some

similarities between random fluctuations. We show two examples in figure 22: total correlation for the left example is 0.242666, and 0.277345 for the right one.

7. Conclusions

To conclude, let us summarize the advantages of the wavelet transform in the context of designing a compact representation of time series for similarity matching.

The Wavelet Transformation incorporates the concept of scale (resolution) to the representation of the time series, which enables us to reveal the scale-wise organisation (hierarchy) of features. Since we are interested in only the largest features, these correspond to events at the largest of the scales of decomposition. The task of selecting such features can be accomplished using the (predefined) number of WT maxima appearing above some largest (predefined) scale of interest.

The Wavelet Transformation provides the possibility of designing a scale-free representation of singular features in time series independent of vertical and horizontal rescaling, shift and invariance with respect to additive (polynomial) bias. In particular, the WT allows us to isolate such singular events and in addition to evaluate their scale free parameters, relative scale, relative position and effective estimate h of their Hölder exponent. We have shown that a set of such features can serve for evaluating the (local) correlation product for time series.

With regard to possible extensions of this methodology, there are two directions which could be taken. One is further to reduce the discrimination power for the sake of rapid rough similarity matching. We have shown that histograms of the local h exponent can be made for all scales. We have also demonstrated that by an appropriate choice of scale or the number of maxima, it possible to incorporate in the distribution the features which belong to the strongest fluctuations and suppress those which belong to the bulk behaviour of the distribution. This approach should be particularly useful in rough matching not oriented at local information.

At the other extreme is the possibility of extending the h-representation over the 'top' scale. This information is not necessary in the case of the very limited scale range covered by the largest features in the h-representation as presented in this report. However, in case a higher discriminating power is needed, more features may have to be included in the representation. In this case, the scale range will likely extend to the degree requiring the top-scale of each singular feature to be incorporated into the representation. Adding a new parameter introduces another degree of freedom in the representation, thus improving the resolution of matching. In particular the scale parameter can be used to implement matching including x rescaling, thus comparing features of different length, or features at different resolutions in one time series.

Due to the character of our application, we have shown how to estimate the Hölder exponent for dense singularities. Of course, there may be applications where isolated singularities are predominant. In this case a more accurate estimation of their exponents can be obtained than by using our multiplicative cascade model. A method to decide whether we deal with an isolated singularity would be useful in this context. The properties of the tree structure revealed by the WT should provide reliable guidance for such a method.

Last but not least, on the implementation side for real data mining applications, compactly supported wavelets from the Haar family will be evaluated for use in efficient and fast algorithms realising the methodology described in this report.

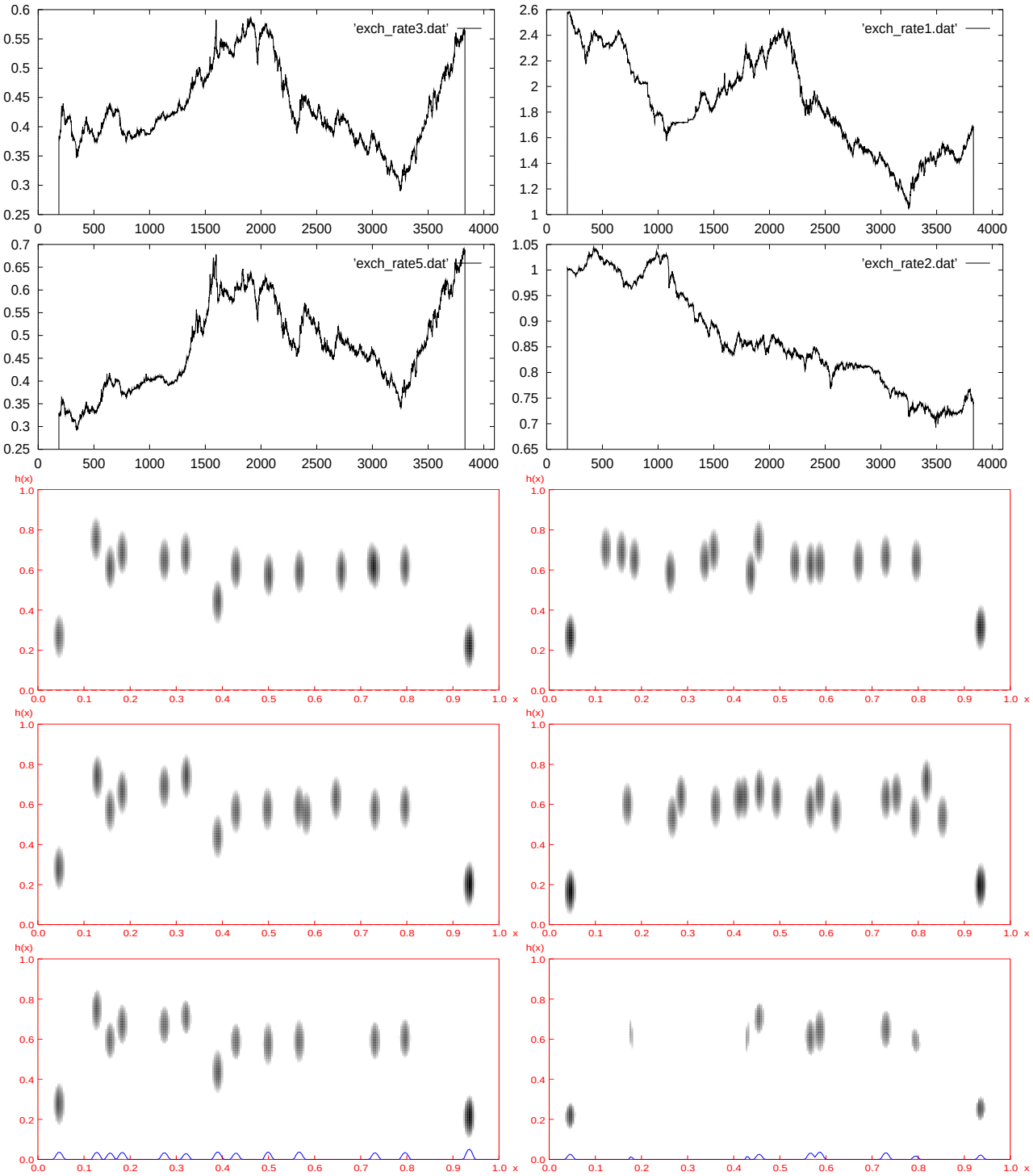

Figure 19: Left: German Mark versus Swiss Franc. Right: Pound Sterling versus Canadian Dollar. The time series in the pair of the upper plots are followed by the distance measures (autocorrelations) obtained from the corresponding h-representations in two plots below. The pointwise correlation of the corresponding h-representations is shown in the bottom plots. Vertical range for pointwise correlation plot is from 0 to 1 for h exponent, horizontal is ranging from 0 to 1 for normalised time. Projection of the pointwise correlation on the time axis is shown with the line plot.

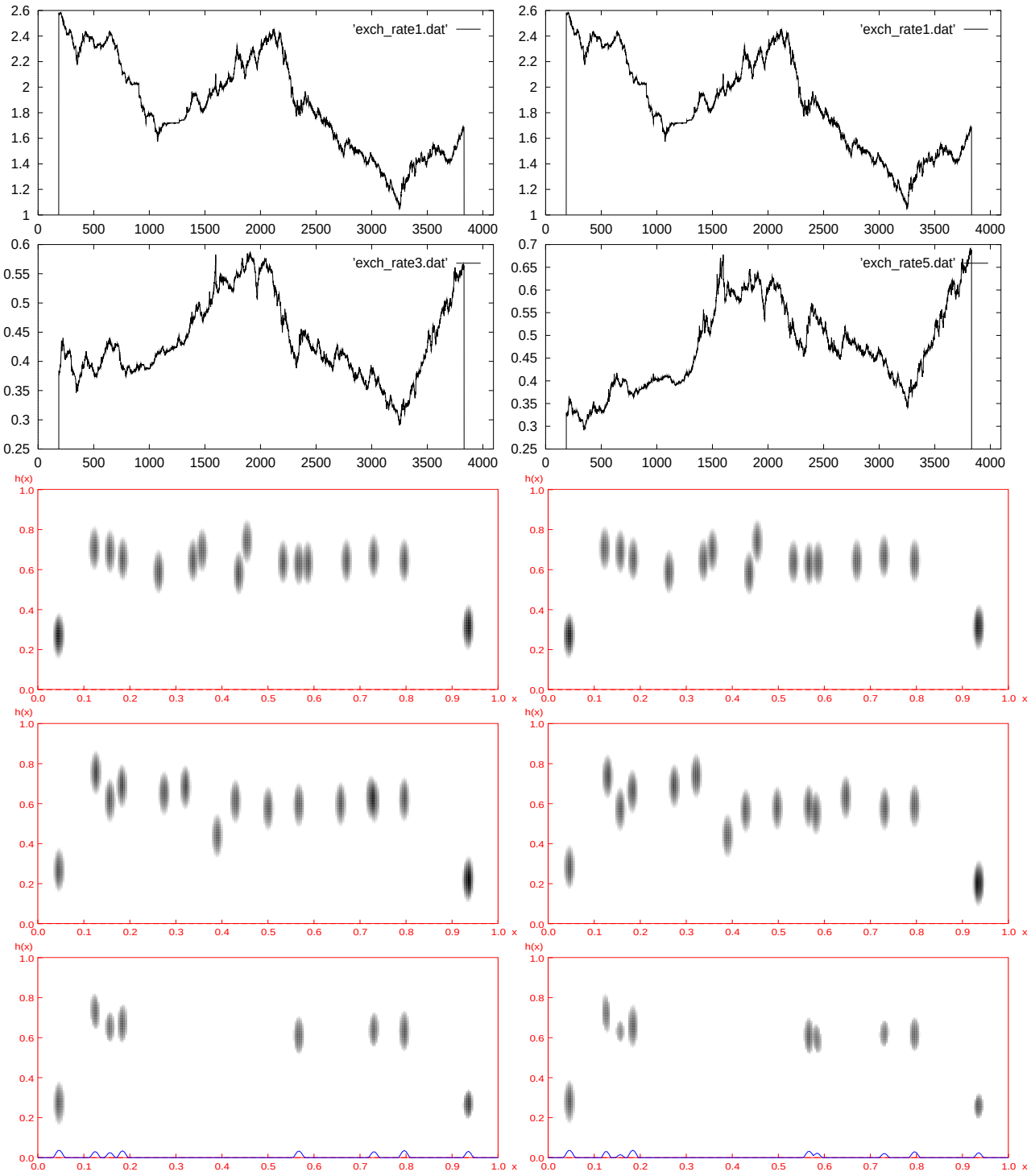

Figure 20: Left: Pound Sterling versus German Mark. Right: Pound Sterling versus Swiss Franc. Description of the axis analogical to that in figure 19.

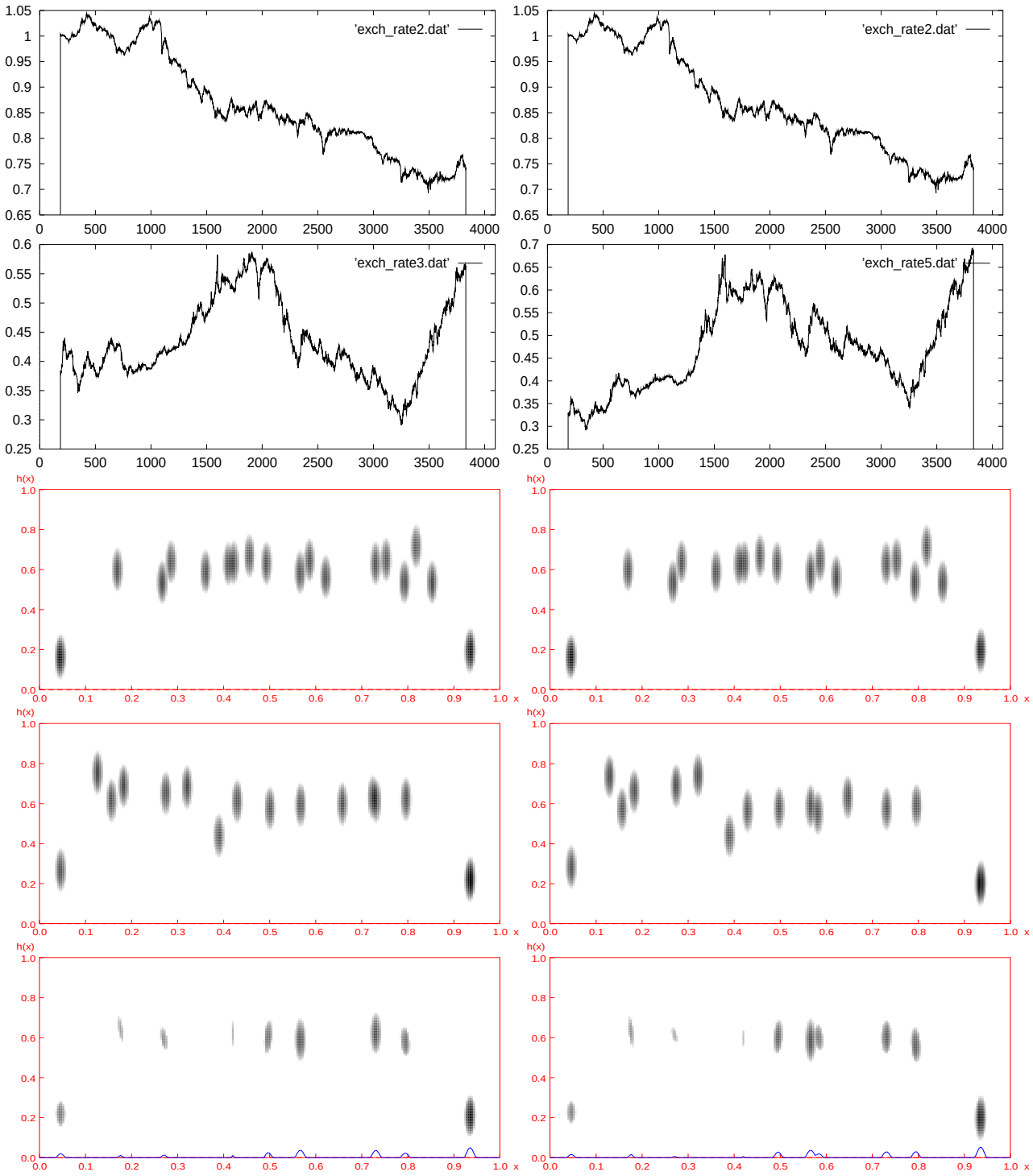

Figure 21: Left: Canadian Dollar versus German Mark. Right: Canadian Dollar versus Swiss Franc. Description of the axis analogical to that in figure 19.

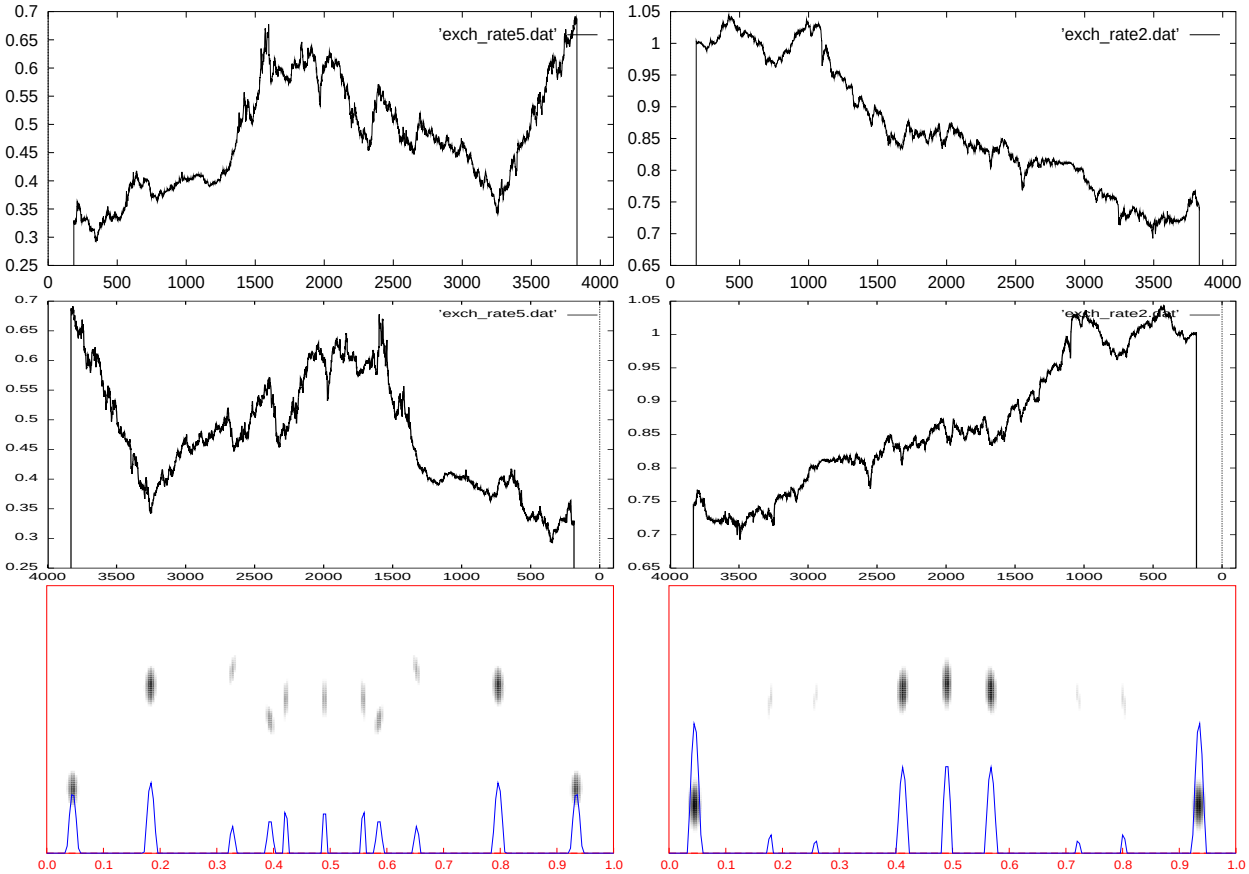

Figure 22: Left: Swiss Franc versus Swiss Franc inverted in time. Right: Canadian Dollar versus Canadian Dollar inverted in time. Description of the axis analogical to that in figure 19.

References

1. Z.R. Struzik, A. Siebes, Wavelet Transform in Similarity Paradigm I, *CWI Report*, **INS-R9802**, (1998), also in *Research and Development in Knowledge Discovery and Data Mining*, Xindong Wu, Ramamohanarao Kotagiri, Kevin B. Korb, Eds, *Lecture Notes in Artificial Intelligence* **1394**, 295-309, Springer (1998).
2. C.J.G. Evertsz, Fractal Geometry of Financial Time Series, *Fractals* **3**, No.3, 609-616 (1995).
3. Y. Liu, P. Cizeau, P. Gopikrishnan, M. Meyer, C.-K. Peng, H.E. Stanley, Volatility Studies of the S&P 500 Index, *preprint*, (1998).
4. I. Daubechies, *Ten Lectures on Wavelets*, S.I.A.M. (1992).
5. M. Holschneider, *Wavelets - An Analysis Tool*, Oxford Science Publications, (1995).
6. A. Arneodo, E. Bacry, S. Jaffard, J.F. Muzy, Oscillating Singularities on Cantor Sets: A Grand-canonical Multifractal Formalism, *preprint*, (1996).
7. S.G. Mallat, W.L. Hwang, Singularity Detection and Processing with Wavelets, *IEEE Trans. on Information Theory* **38**, 617-643 (1992).
8. S.G. Mallat, S. Zhong, Complete Signal Representation with Multiscale Edges, *IEEE Trans. PAMI* **14**, 710-732 (1992).
9. Z.R. Struzik, The Wavelet Transform in the Solution to the Inverse Fractal Problem, *Fractals* **3** No.2, 329-350 (1995).
10. Z.R. Struzik, Removing Divergences in the Negative Moments of the Multi-Fractal Partition Function with the Wavelet Transformation, *CWI Report*, **INS-R9803**, (1998).
    Also in 'Fractals and Beyond - Complexities in the Sciences', M.M. Novak, Ed., World Scientific, 351-352, (1998).
11. K. Falconer, *Fractal Geometry - Mathematical Foundations and Applications*, John Wiley (1990).
12. Z.R. Struzik, Fractals under the Microscope or Reaching Beyond the Dimensional Formalism of Fractals with the Wavelet Transform, *CWI Quarterly*, **10**, No 2, 109-151 (1997).
13. A. Arneodo, E. Bacry, J.F. Muzy, The Thermodynamics of Fractals Revisited with Wavelets, *Physica A*, **213**, 232-275, (1995).
14. J.F. Muzy, E. Bacry, A. Arneodo, The Multifractal Formalism Revisited with Wavelets, *International Journal of Bifurcation and Chaos* **4**, No 2, 245-302 (1994).