Measuring Time Series' Similarity through Large Singular Features Revealed with Wavelet Transformation

Zbigniew R. Struzik, Arno Siebes

Kruislaan 413, 1098 SJ 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 task of comparing records of time series. 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 addition, we are able to reveal scale-wise ranking of singular events including their scale free characteristic: the Hölder exponent. We use a set of 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 with 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.

Note: This work has been carried out under the Impact project.

1. Introduction

The importance of similarity measures in data mining is easily underestimated. This is caused by the fact that most algorithms assume relational data and the similarity is implicitly measured by similarity (or even equality) of values for given attributes. However, the importance of similarity measures becomes apparent the moment one considers mining non-relational data, such as time series. In such cases, patterns should describe sets of time series with similar behaviour and, thus, a similarity measure is necessary.

This similarity measure could be implicit in completely new algorithms which work for this special type of data. Or, the similarity measure could be explicit and use the standard data mining algorithms. In the context of time series, one way to achieve this is to extract a (fixed) number of features from the time series so that similar time series have similar features (e.g., as numerical values) and vice-versa. This is the approach we, and others [1, 2, 3, 4, 5], follow in our research. In other words, we want to define the similarity of time series through a number of features.

In general, 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. On the other hand, the subjective, qualitative judgement of similarity (by humans) is based precisely on non-stationary behaviour; rapid transients marking beginnings of trends, extreme fluctuations and generally speaking, large but rare events.

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 a statistical point of view can be of advantage in another context. In this paper, 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 this paper, we will demonstrate how the Wavelet Transform resulting from scale-wise decomposing of time series data provides a natural way 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 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 between such singularities and 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 3, 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 in section 4. This is followed by the test case of correlating examples of currency exchange rates. Section 5 closes the paper with conclusions and suggestions for future developments.

2. 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. [6, 7], provides a way of analysing the 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. One of the aspects of the WT which is of great advantage for our purpose is the ability to reveal the *hierarchy* of (singular) features including their scaling behaviour - the so-called *scale free* behaviour.

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 paper, we have chosen 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:

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

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

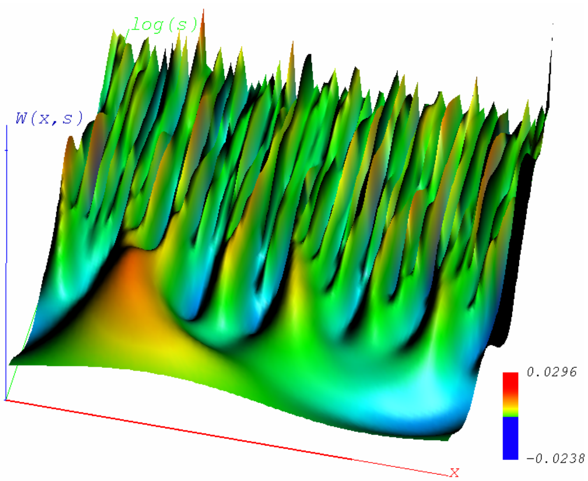

Figure 1. Continuous Wavelet Transform representation of the random walk (Brownian process) time series. The wavelet used is the Mexican hat - the second derivative of the Gaussian kernel. The coordinate axis are: position x, scale in logarithm $\log(s)$, and the value of the transform $Wf(s,b)$.

In figure 1, 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 all the temporal locality properties - the position axis is in the forefront of the 3D plot. 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.

A useful representation which can be derived from the CWT and which is 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 local singular behaviour of time series.

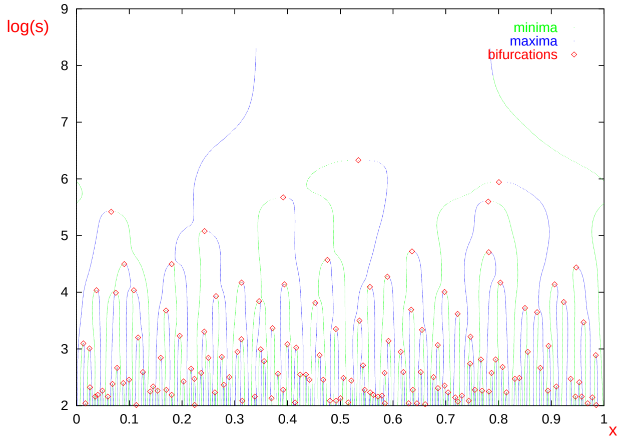

Figure 2. WTMM representation of the time series and the bifurcations of the WTMM tree. Mexican hat wavelet.

It consists of the maxima lines extracted from the CWT, see figure 2. Each line is constructed from local maxima of the CWT with respect to the time coordinate, connected along the scale. It can be shown that each such a maximum line converges to a singularity in the time series [1], thus making possible the localisation of the singularity. Moreover, it can be used for the evaluation of the Hölder exponent of the singularity:

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

if $h(x_0) < n + 1$, where n indicates the number of the vanishing moments of the wavelet (in the order of the derivative of Gaussian kernel $d^n\theta(x)/dx^n$ in our case). The meaning of the Hölder exponent can be loosely associated with the feeling of local roughness or regularity of the time series. The higher the Hölder exponent, the more smooth and regular the time series in x_0.

Suppose the time series can be locally approximated with the Taylor series expansion of f around x_0 up to the order n:

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

Of course, if $h(x_0)$ is smoother than some n-th degree polynomial, the polynomial bias has to be removed in order to access the singular behaviour $h(x_0)$. This is why we are using a wavelet with n vanishing moments, which effectively filters the n-th degree polynomials bias in the time series.

3. The h-Representation

As already discussed in section 2, 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.

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. The h-representation, as we will call it, will be obtained by means of selecting a predefined number of strongest maxima and then tracing them *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. [2]

For the sake of comparison, we plot in figure 3 left, the sampling of the input time series with 10 and 25 maxima first appearing while going down from the highest scale - lowest resolution. There is a substantial amount of detail added to the 'approximation' with 25 maxima compared to that with 10, nevertheless the strongest features remain unchanged. In figure 3 right, we compare the sampling with the 10 strongest maxima against the original time series. Again, the largest features are well captured by the sampling proposed.

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 are taken into the representation. The information carried by the sign is complementary to that in the Hölder exponent - the pair ($h(x_i)$,sign($Wf(x_i)$)) is somewhat reminiscent of the (amplitude, phase) in the Fourier Transform. In particular, the sign will allow us to distinguish the time series $f(x)$ and its inverted version $-f(x)$.

[1] on condition that the number of vanishing moments of the wavelet is sufficient.

[2] For practical details on h-exponent estimation see [9, 10].

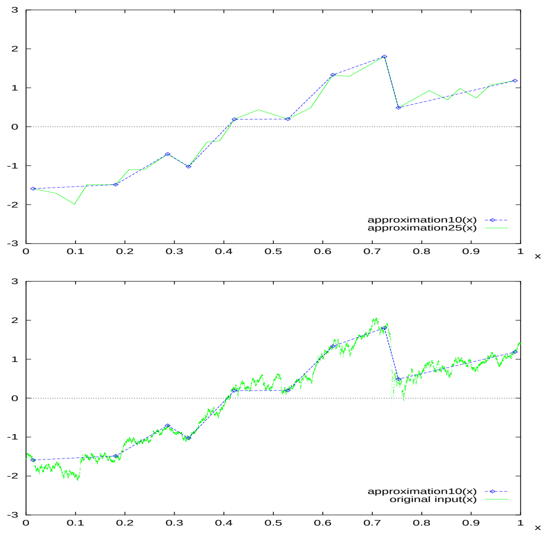

Figure 3. Top: the ‘approximation’ of the time series using the 10 strongest maxima, overlayed onto the ‘approximation’ using all 25 maxima at the scale considered $\log(s) = 6$. Bottom: the ‘approximation’ of the time series using the strongest 10 maxima, overlayed onto the original time series.

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 $h(x_i)$, and the corresponding sign of the wavelet transform $Wf(x_i)$.

4. Experiments with Similarity

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:

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

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:

$$corr_{s_1,s_2}(x,h) = dist_{s_1}(x,h)\, dist_{s_2}(x,h)\ . \tag{2}$$

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 the missing values.) Below, in figure 4, we show the plots of the records.

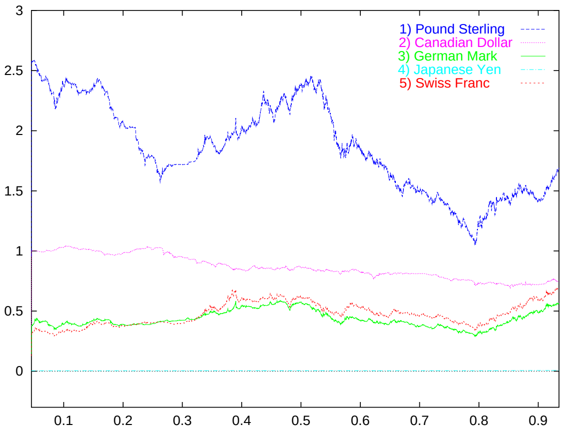

Figure 4. Left above, all the records of the exchange rate used, with respect to USD over the period 01/06/73 - 21/05/87.

All the time series were decomposed using the Mexican hat wavelet. For each, the $15 - 20$ 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: [3]

1) German Mark(s_3) versus Swiss Franc(s_4); total correlation = 0.793370

2) Pound Sterling(s_1) versus Canadian Dollar(s_2); total correlation = 0.287755

3) Pound Sterling(s_1) versus German Mark(s_3); total correlation = 0.408833

4) Pound Sterling(s_1) versus Swiss Franc(s_4); total correlation = 0.375356

5) Canadian Dollar(s_2) versus German Mark(s_3); total correlation = 0.314108

6) Canadian Dollar(s_2) versus Swiss Franc(s_4); total correlation = 0.337519.

[3] 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.

The pointwise correlations of the corresponding h-representations for the two example pairs are shown in figure 5. Even at the very low resolution of the h-representations, the correlation plot conveys relevant temporal information about the local similarity of time series matched. The time series s_3 and s_4 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.

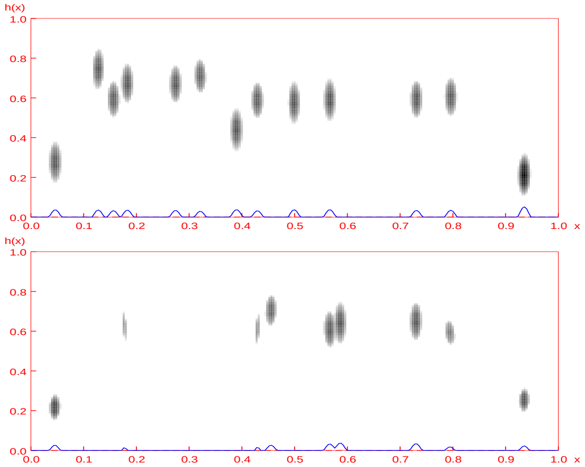

Figure 5. Top: German Mark versus Swiss Franc. Bottom: Pound Sterling versus Canadian Dollar. The pointwise correlation of the corresponding h-representations is shown in the bottom plots.

A possible interpretation is that the time series s_3 and s_4 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.

5. Conclusions

We have demonstrated that through incorporating the concept of scale (resolution) to the representation of the time series, the Wavelet Transformation 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 WT allows us to evaluate the scale free parameters of the isolated singular events: relative scale, relative position, sign and the Hölder exponent. We have shown that a set of such features can serve for evaluating the (local) correlation product for time series.

References

[1] R. Agrawal, K-I. Lin, H.S. Sawhney, K, Shim, Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time Series Databases, in *Proceedings of the 21 VLDB Conference*, Zürich, 1995.

[2] R. Agrawal, C. Faloutsos, A. Swami, Efficient Similarity Search in Sequence Databases, in *Proc. of the Fourth International Conference on Foundations of Data Organization and Algorithms*, Chicago, 1993.

[3] G. Das, D. Gunopulos, H. Mannila, Finding Similar Time Series, in *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial intelligence 1263, Springer, 1997.

[4] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, Fast Subsequence Matching in Time-Series Databases", in *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 1994.

[5] Z.R. Struzik, A. Siebes, Wavelet Transform in Similarity Paradigm, in *Research and Development in Artificial Intelligence*, Xindong Wu, Ramamohanarao Kotagiri, Kevin B. Korb, Eds, *Lecture Notes in Artificial Intelligence* **1394**, Springer, (1998).

[6] I. Daubechies, *Ten Lectures on Wavelets*, S.I.A.M. (1992).

[7] M. Holschneider, *Wavelets - An Analysis Tool*, Oxford Science Publications, (1995).

[8] S.G. Mallat, S. Zhong, Complete Signal Representation with Multiscale Edges, *IEEE Trans. PAMI* **14**, 710-732 (1992).

[9] Z.R. Struzik, A. Siebes, Wavelet Transform in Similarity Paradigm II, *CWI Report*, **INS-R9815**, CWI, Amsterdam (1998).

[10] Z.R. Struzik, Local Effective Hölder Exponent Estimation on the Wavelet Transform Maxima Tree, in *Fractals: Theory and Applications in Engineering*, Michel Dekking, Jacques Lévy Véhel, Evelyne Lutton, Claude Tricot, Eds. Springer Verlag (1999).