Calculating Hurst exponent and fractal dimension of neutron monitor data in a single parallel algorithm

A S Kussainov¹ and S G Kussainov²

¹ Physics and Technology Department, al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan
² Institute of Industrial Engineering, K.I. Satpaev Kazakh National Technical University, Almaty, Republic of Kazakhstan
E-mail: arman.kussainov@gmail.com

Abstract.

We implemented an algorithm for simultaneous parallel calculation of the Hurst exponent $H$ and the fractal dimension $D$ for the time series of interest. Parallel programming environment was provided by OpenMPI library installed on three machines networked in the virtual cluster and operated by Debian Wheeze operating system. We applied our program for a comparative analysis of week and a half long, one minute resolution, six channels data from neutron monitor. To ensure a faultless functioning of the written code we applied it to analysis of the random Gaussian noise signal and time series with manually introduced self-affinity features. Both of them have the well-known values of $H$ and $D$. All results are in good correspondence with each other and supported by the modern theories on signal processing thus confirming the validity of the implemented algorithms. Our code could be used as a standalone tool for the different time series data analysis as well as for the further work on development and optimization of the parallel algorithms for the time series parameters calculations.

1. Introduction

The applications of the Hurst exponent $H$ and fractal dimension $D$ calculations are ranging from stock market analysis [1] to electron gas modeling [2] addressing data statistics and systems fractal properties [3]. Multiple studies have been done, including the studies of abundant data on cosmic rays variations. For example, Patra et al [4] analyzed 36 years long data series on cosmic rays density covering time period almost three solar cycles long and came to the conclusion that “the present data is anti-persistent in behavior and the process is a short memory process with the $H$ value of 0.15”. The data size does not always guarantee or necessary to perform an informative study of the subject. Flynn and Pereira, on the contrary, studied extra short, hundred points and less, data sequences [5] and extracted vital information from a data sample on population dynamics.

A fractal dimension $D$ is a quantity reflecting how frequently and to what extent the self-affine patterns appear with the change of scale in picture. For self-affine processes we expect the local properties to be observable at the level of the global ones leading, in limited number of cases, to the relationship $D + H = n + 1$ between $D$ and $H$ in $n$-dimensional space [6]. More often, in real experimental data, local and global behaviors are decoupled without any linear relation.
Hurst exponent and fractal dimension estimates are mainly dependent on the nature of data, the length of data sample, the signal to noise ratio and the range of scales used for the analysis.

Our primary goal in this paper is to perform an effective parsing of original data text file, to identify similar steps in $H$ and $D$ calculations algorithms, and to provide reliability of results of the Hurst exponent $H$ and fractal dimension $D$ data calculations. We also demonstrate a quick self-affinity test for the data of interest using our program.

2. Methods and results

Conventional algorithm for the Hurst exponent calculation runs as follows: Original time series of length $N$ are divided into the sets of shorter partial series $\{X_i\}$ with length $n = N, N/2, N/4, \ldots, 4, 3, 2$ points. The upper, $n = N$, and lower, $n = 2$, cutoff limits differ from study to study and depend on data availability and studied phenomena. For each set with particular $n$ value, and for every partial series $\{X_i\}$ within this set, we calculate the mean-adjusted series $\{Y_i\}$ derived from $\{X_i\}$ by subtracting from each element the series' mean value $m$

$$Y_t = X_t - m; \quad t = 1, 2, \ldots, n; \quad m = 1/n \sum_{i=1}^{n} X_i$$  \hspace{1cm} (1)

We define the rescaled range $R(n)$ for a given subseries $\{Y_i\}$ as follows

$$R(n) = \max(Z_1, Z_1, Z_n) - \min(Z_1, Z_1, Z_n)$$  \hspace{1cm} (2)

where the cumulative deviate series $Z_n$ are given by the following equation

$$Z_t = \sum_{i=1}^{t} Y_i; \quad t = 1, 2, \ldots, n.$$  \hspace{1cm} (3)

Now, we can built the tabulated power law function $E(n)$ for all possible values of $n$ such as

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^H$$  \hspace{1cm} (4)

where $S(n) = \{1/n \sum_{i=1}^{n} (X_i - m)^2\}^{1/2}$ is a standard deviation. The value of $H$ then could be estimated from the slope of the line $\log(E) = \log(C) + H \cdot \log(n)$ [7].

Fractal dimension $D$ is closely related to the Hurst exponent and could be calculated, according to Higuchi cornerstone paper [3], by constructing the following sets of subseries $X_k^n$

$$X(m), X(m + k), X(m + 2k), \ldots, X(m + \left[\frac{N - m}{k}\right] k);$$
$$m = 1, 2, \ldots, k; \quad k = 1, 2, \ldots, \left[2^{(j-1)/4}\right]; \quad j = 11, 12, 13, \ldots$$  \hspace{1cm} (5)

where the square brackets are used to denote the closest integer after rounding the fraction to zero.

Next, we calculate the normalized lengths $L_m(k)$ of the constructed subseries

$$L_m(k) = \left\{\frac{N - 1}{\left[\frac{N-m}{k}\right]} k \sum_{i=1}^{\left[\frac{N-m}{k}\right]} |X(m + ik) - X(m + (i-1)k)|\right\} k^{-1}$$  \hspace{1cm} (6)

These are expected to follow the similar power law in the form of $\langle L_m(k) \rangle \propto k^{-D}$ after averaging within all sets of $k$ values for different $m = 1, 2, \ldots, k$. 

\hspace{1cm}
Table 1. Curves slope estimates in $D$ and $H$ calculations.

| Time series, name       | Time series length, $N$ | Fractal dimension, $D$ | Hurst exponent, $H$ |
|-------------------------|-------------------------|------------------------|---------------------|
| Higuchi series          | $2^{17}$                | -1.5345                | 0.9875              |
| Gaussian noise          | $1000 + 2^{17}$         | -1.9999                | 0.5400              |
| NM data, 1st channel    | 14703                   | -1.8128                | 0.9810              |
| 2nd channel             | 14703                   | -1.9015                | 0.9332              |
| 3rd channel             | 14703                   | -1.9363                | 0.9184              |
| 4th channel             | 14703                   | -1.9583                | 0.8734              |
| 5th channel             | 14703                   | -1.9064                | 0.9333              |
| 6th channel             | 14703                   | -1.9102                | 0.9425              |

To address a persistent need for the effective parallel algorithms in data processing we designed our piece of C/C++ code compatible with available for all latest Debian operating system distribution of the Message Passing Interface (OpenMPI) parallel computations environment. We have configured our virtual cluster using the Oracle VM VirtualBox. The host is 64 bit Windows 8.1 operating system running on Intel Core i3-3220 CPU with 4 Gb of RAM. The guest operating systems are the three Linux machines with 64 bit Debian Wheezy 7.4 with 512 Mb of RAM per each server and two nodes. The number of processes was equal to the number of virtual machines and it was arranged to be three. All scales calculations were distributed between them sequentially in a circular fashion, see Figure 2.

The original neutron monitor data were retrieved from the underground monitor of the Tien Shan High-Altitude Station (3340 m above sea level) located in Almaty, Kazakhstan. The monitor has several sections with multiple detectors located under a ground layer 20 m of water equivalent thick [8]. Neutron counts were acquired at one minute interval from January the 1st, 2015 till January the 11th, 2015. All data are pressure corrected. Neutron monitor time series data additionally underwent simple exponential smoothing filtration procedure as described in [9].

Our $H$ estimate for the first out of six registration channels of the nuclear monitor data file is shown on Figure 1. Each channel data contain week and a half long data on neutrons counts, taken with one minute resolution. Operator makes the choice of registration channel for analysis within the program. Calculated $H$ and $D$ values for all of the channels are given in Table 1.

To test the validity of the written program we reproduced the exact same time series $Y(i)$ (we using the original $Y(i)$ and $Z(j)$ notations from his paper) which were used by Higuchi in his derivations [3] and plotted the results on the same figure for reference, see Figure 1. Here $Y(i) = \sum_{j=1}^{1000+i} Z(j)$, where $Z(j)$ is a Gaussian noise with mean zero and standard deviation equals to 1. As we said, the set of $\langle L_m(k) \rangle$ values is expected to represent the power law function $\langle L_m(k) \rangle \propto k^{-D}$ and the value of $D$ could be extracted within the same piece of code we wrote previously for the Hurst exponent.

Our calculations give us $D$ value equals $-1.5345$ and $H = 0.9875$ for the Higuchi’s time series. A separate test run of the code for the Gaussian time series $Z(j)$ produced another set of values $H$ and $D$ equal $-1.9999$ and $0.5400$ correspondingly, see Table 1. These values are typical for the highly uncorrelated times series. Both sets of scales $n$ and $k$ are taking the same values ranging from 6 to 6888. The $Y(i)$, $Z(j)$ original series and neutron monitor data are plotted as the offsets on the same Figure 1.

The $\pm \Delta y_i$ error bars from the least squares fitting procedures were plotted on the Figure 1.
around each experimental data points \((x_i, y_i)\) using \texttt{polyfit}, \texttt{polyval} and \texttt{errbar} functions available in GNU \texttt{Octave} software in such a way that then \(y_i \pm \Delta y_i\) contains at least 50% of the predictions of future observations at \(x_i\), see \texttt{Octave} help notes on the selected functions.

**Figure 1.** Calculations of \(H\) and \(D\) values for the Higuchi time series, the neutron monitor data and the Gaussian noise. Offsets on the right side contain the plots of analyzed time series.

As we mentioned before, the \(n\)-length of subseries for the Hurst exponent calculations, and \(k\)-value for the fractal dimension calculation, have been treated equally, as a single variable \(k = n\), in our code. That is two separate arrays of the same size \(k\) were allocated for both types of concurrent calculations of the \(H\) and \(D\) values. In case of the Hurst exponent calculations the array’s values have been filled sequentially from the data file readout. For the fractal dimension calculations, each array’s element contained the single value \(\langle L_m(k) \rangle\) was updated through the same data readout according to the scheme described in equations (5) and (6).

Calculated slopes for all three sets of data, where neutron monitor data contain information on all six channels, are given in the Table 1. These are the Higuchi time series, Gaussian noise
Figure 2. Typical workload distribution between the three processes on the virtual cluster with timing output.

and neutron monitor data $H$ and $D$ values. With such an approach, the $H$ and $D$ values may be calculated retrospectively, for the whole history of data acquisition, as well as at the present moment, within a fixed time window moving forward with the natural time flow and incorporated with the acquisition hardware.

3. Discussion

$H$ and $D$ values for the Higuchi time series in our calculations were found to be equal to the $D$ and $H$ values from [3], that is $H$ value is close to 1 (see Table 1) and indicates long term positive autocorrelation, as expected from the $Y(i)$ series. For the neutron monitor series, Hurst exponent estimates are matching the majority of the previously obtained results for geomagnetic indices [10] where $H$ value is well above 0.5. Drastic changes in $H$ and $D$ are observed if rigorous filtering is applied to the raw data. This is usually the case with intrinsically very noisy neutron monitor data. Between all six channels, the $H$ values are within $0.9333 - 0.9810$ range, suggesting that at the chosen series duration and quality, our time series do have some scalable order. However, possibility of a long-term positive autocorrelation requires further studies and clarification.

The Gaussian noise data $Z(j)$ in its turn produces values of $D=-1.9999$ and $H=0.5400$ as expected from the highly uncorrelated data. As we can see, for all sets of data the processes are not completely self-affine in the sense of $D + H = n + 1$ relationship, and the dependence between $D$ and $H$ is not of linear nature.

It is interesting to observe how the curve's slope changes according to the changes in data structure. We can see this transition when we come down from Higuchi time series to the chaos in the Gaussian noise. All lines for the Hurst exponent originate from the single point. That is rotations for all $H$ data happen to be around $(2,0)$ point due to the loss of differences between time series at the small scales. Through such a graphic representation and comparative, real-time analysis of our data with the purpose of data quality check and search for a particular fractal dimensional phenomena is possible.
4. Conclusions
Using C/C++ programming language we implemented an algorithm for the simultaneous parallel calculations of the Hurst exponent $H$ and the fractal dimension $D$ for the different time series. Parallel programming environment was provided by an open source freeware such as OpenMPI package installed on three machines networked to the virtual cluster using the Oracle VM VirtualBox and operated by a 64 bit Debian Wheeze 7.4 operating systems. Additional data processing and graphic representation were made on GNU Octave freeware.

We used our program to perform a comparative analysis of the week and a half long, one minute resolution, six channels data file from a neutron monitor. To verify the validity of the written code, we compared these results with a similar data analysis of the random Gaussian noise signal and with time series which have manually introduced self-affinity features. Both reference time series have the known values of $H$ and $D$. All results are in good correlation with each other and supported by the modern theories on signal processing, thus confirming the validity of the implemented algorithms.

Our algorithm has multiple applications such as quick data self-affinity [11] test and have great potentials for the future development. The code could be used as a tool to study time series of different nature and origin. Timing and optimization in this paper are the subjects for the further studies.

Acknowledgments
This study was supported by grant #2532/GF3 provided by the Science Committee at the Ministry of Science and Education of Republic of Kazakhstan to the principal investigator at the National Nanotechnology Open Laboratory. We also would like to thank Dr. Yelena White, from East Georgia State College, Swainsboro, GA, USA, for editing the manuscript.

References
[1] Alvarez-Ramirez J, Alvarez J, Rodriguez E and Fernandez-Anaya G 2008 Physica A 387 6159-69
[2] Hurst J, Morandi O, Manfredi G and Hervieux P 2014 EPJ D 68 176-179
[3] Higuchi T 1988 Physica D 31(2) 277-283
[4] Patra S N, Ghosh K and Panja S C 2013 IJEER 3(2) 237-246
[5] Flynn M N, Pereira W R L S 2013 Biomatematica 23 1-14
[6] Mandelbrot B B, Passoja D E and Paullay A J 1984 Nature 308 721-724
[7] Górski A Z, Drożdż S and Speth J 2002 Physica A 316 496-510
[8] Shepetov Yu V 2014 The Tien-Shan mountain station’s database (Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation named after Nikolay Pushkov of the Russian Academy of Sciences) http://cr29.izmiran.ru
[9] Brown R G 1956 Exponential smoothing for predicting demand (Cambridge, Massachusetts: Arthur D. Little Inc) p 15
[10] Pesnell W D 2012 Sol. Phys 281 507532
[11] Gneiting T and Schlather M 2004 SIAM Rev. 46(2) 269-282