RETURN INTERVAL DISTRIBUTION OF EXTREME EVENTS IN LONG MEMORY TIME SERIES WITH TWO DIFFERENT SCALING EXPONENTS

A thesis submitted towards partial fulfillment of

BS-MS dual degree program

by

Smrati Kumar Katiyar

(20061015)

Under the guidance of

Dr. M S Santhanam

(Assistant Professor,IISER pune)

to the

DEPARTMENT OF PHYSICS

INDIAN INSTITUTE OF

SCIENCE EDUCATION AND RESEARCH

PUNE - 411008

APRIL 2011

CERTIFICATE

This is to certify that this dissertation entitled "**RETURN INTERVAL DISTRIBUTION OF EXTREME EVENTS IN LONG MEMORY TIME SERIES WITH TWO DIFFERENT SCALING EXPONENTS** towards the partial fulfillment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research Pune, represents original research carried out by **Smrati Kumar Katiyar**(20061015), at IISER, pune under the supervision of **Dr. M S Santhanam**(Assistant Professor) during the academic year 2010-2011.

Smrati Kumar Katiyar

Supervisor	Head
(Dr. M S Santhanam)	(Department of Physics)
Date:	Date:
place:	place:

Acknowledgement

I would like to thank my project guide **Dr. M S Santhanam** for his continuous help in my project He always listened to me and gave his continuous advice to clarify my doubts.

Abstract

Many processes in nature and society show long memory. In recent years research towards the return intervals of extreme events has picked a significant pace due to its practical application in several different fields ranging from geophysics,medical sciences,computer science to finance and economics. Earlier results on the return interval distribution of extreme events for long memory process with single scaling exponent have shown that the distribution is a product of power law and stretched exponential function. In this report, we have obtained an analytical expression for the return interval distribution of long memory processes with two different scaling exponents. We also provide numerical simulations to support our analytical results.

Contents

Chapter 1

Introduction

Study of extreme events is an active area of research presently for its inherent scientific value and its practical application to diverse fields ranging from medical sciences, geophysics, computer science to economics and finance[1][2]. The consequences of extreme events such as earth quakes or cyclones are generally catastrophic to human society. Hence the study of extreme events assume special significance. For instance, in the field of geoscience it could help us in the prediction of next earthquake. In the field of finance these kind of studies can help us to make better prediction about market crashes. Several natural and socio-economic phenomenon such as atmospheric pressure, temperature, volatility etc. show long term memory and hence it is relevant to study the return interval distribution of extreme events for long memory processes. Time series which show power law type autocorrelation function[1][3] with very slow decay in correlation for large lags, as compared to other uncorrelated processes, generally represent long memory processes.

[1] Autocorrelation function(ρ_k) for a given time series $x(t)$ at lag k is defined as $\rho_k = \frac{\sum_{t=k+1}^{n}(x_t-\overline{x})(x_{t-k}-\overline{x})}{\sum_{t=1}^{n}(x_t-\overline{x})^2}$ for $k = 1, 2, ..$

Let $x(t)$ be a given time series such that $\langle x \rangle = 0$, all those events for which $x(t) > q$ will be called extreme events. q is the threshold value.

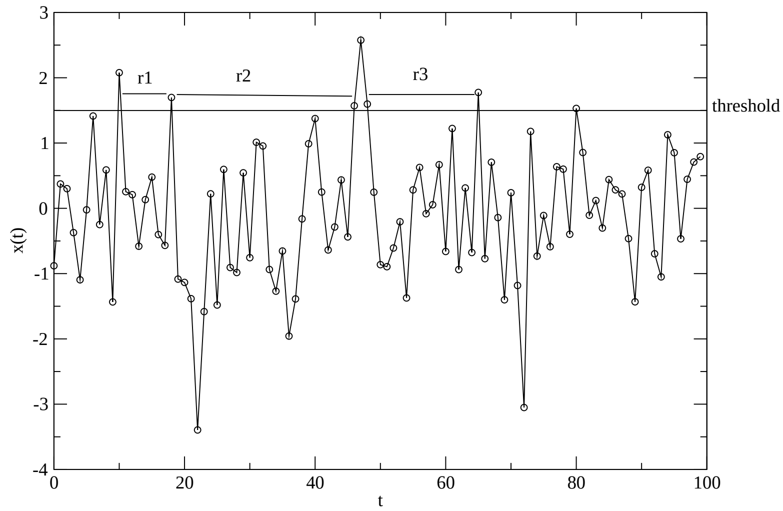


Figure 1.1: *a schematic diagram shows the return intervals for a threshold value q=1.5 as a function of time t*

The return interval r is defined as the time interval between the consecutive extreme events. So we will get a series of return intervals r_k $, k = 1, 2, 3, ..., N$

For uncorrelated processes the return interval distribution is[1]

$$P(R) = e^{-R}. \tag{1.1}$$

For long memory process autocorrelation function is $\rho_k \sim k^{-\gamma}$, where $0 < \gamma < 1$ and return interval distribution of extreme events is [8]

$$P(r) = ar^{-(1-\gamma)}e^{-(\frac{a}{\gamma})r^{\gamma}}. \tag{1.2}$$

In this a is a constant to be determined from normalisation condition. The result in Eq. 1.2 is obtained for a long memory time series with one scaling exponent. What happens to time series with more than one scaling exponent (one of the example for such time series is *high frequency financial data*) ? Figure 1.2 shows one of such case.

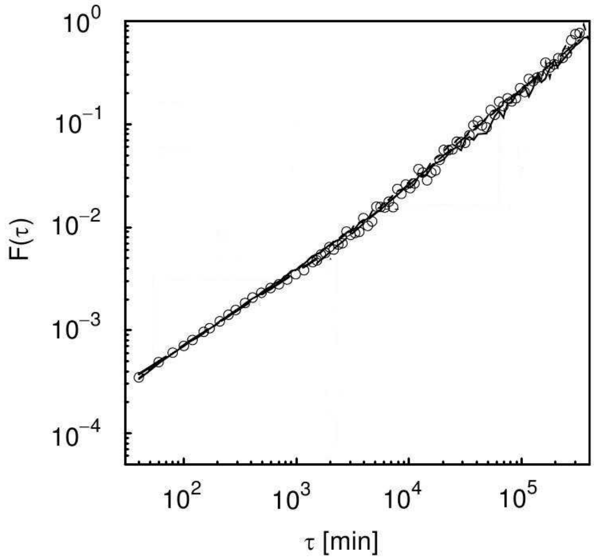
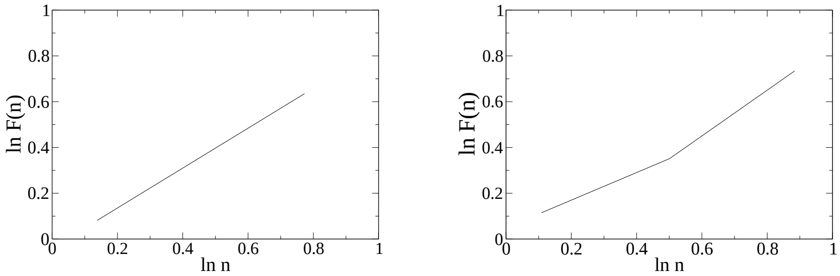


Figure 1.2: *Log log plot of fluctuation function $F(\tau)$ of log returns for the high-frequency data of the S and P 500 stock index, circles represent actual data, line represents linear fitting. (figure taken from Podobnik et al*[10])

Figure 1.3(a) and 1.3(b) shows a comparison between time series with one scaling exponent and two scaling exponents based on fluctuation analysis (Ex. DFA, R/S analysis)[4][5].In these figures x-axis represents time scale n and y-axis $F(n)$ represents averaged fluctuation function. The core of this report will address this problem and provide analytical and numerical solutions to it. In chapter 2, we will present a detailed study of a model for long memory



(a) One scaling exponent (b) Two scaling exponent

Figure 1.3: *Fluctuation analysis of time series*

process. Using a variant of this model, we generate time series with long memory and two time scales. Chapter 3 will be dedicated to the analytical solution for return interval distribution while in chapter 4 we will present a numerical solution. In chapter 5 we will present long memory probability process whose return intervals are uncorrelated and compare the analytical results with the numerically simulated data. In chapter 6 we will discuss our analytical and numerical results and also discuss some practical limitations of our results.

Chapter 2

Long memory process

The plot of sample autocorrelation function (ACF) (ρ_k) against lag k is one of the most useful tool to analyse a given time series. If the ACF tends to zero after some lag k, then the traditional stationary ARMA process[3] are good enough to describe the time series. However if the ACF decays slowly, it means that even after very large lags the value of ACF is sufficiently large and describing the time series with traditional ARMA models will result in excessive number of parameters.

So how we can describe these kind of processes ?

The answer is "*long memory process* ". Formally, "a stationary process will have long memory, if its autocorrelation function can be represented as

$$\rho_k \to C_\rho k^{-\gamma} \quad as \;\; k \to \infty \tag{2.1}$$

$$\text{where } C_\rho > 0 \;\; and \;\; \gamma \in (0, 1)$$

ARFIMA process *introduced by Granger and Joyex in 1980* are known to be capable of modelling long memory processes. The general form of

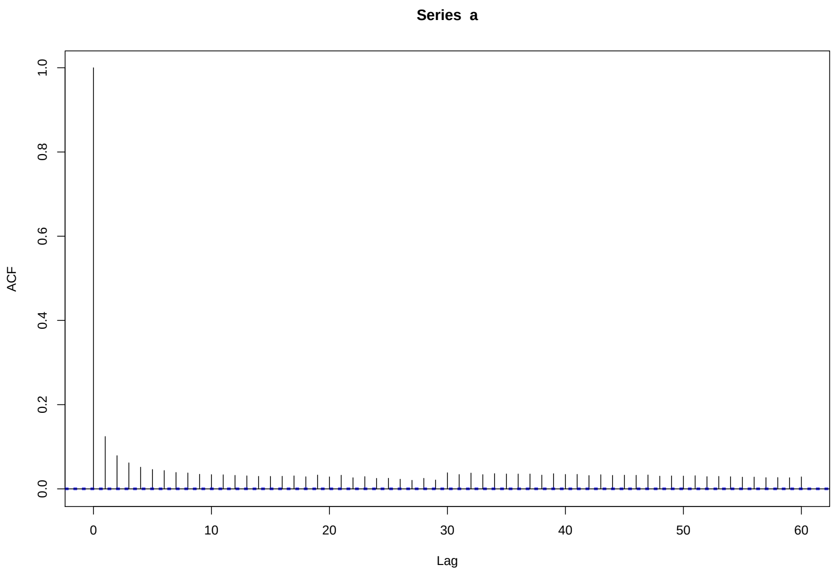


Figure 2.1: *Autocorrelation function for a long memory process*

ARFIMA model of order (p, d, q) is [7]

$$\phi(B)(1-B)^d x_t = \theta(b)a_t. \tag{2.2}$$

This equation can also be written as

$$(1-B)^d x_t = \frac{\theta(B)a_t}{\phi(B)}.$$

In equation(2.2), d is the differencing parameter and it is a fraction, x_t is the original time series, B is the backshift operator defined as $Bx_t = x_{t-1}$, $B^2 x_t = x_{t-2}$ and so on..

In this, $\phi(B)$ is the autoregressive operator

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ....\phi_p B^p,$$

$\theta(B)$ is the moving average operator

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ...\theta_q B^q,$$

and a_t is the white noise with zero mean and variance σ^2.

$$a_t \sim WN(0, \sigma^2).$$

Now for further discussion of ARFIMA process we will consider that p and q are both zero. So the Eq. (2.2) will become

$$(1-B)^d x_t = a_t, \tag{2.3}$$

which is known as fractionally differenced white noise (FDWN). Equation(2.3) can also be written as

$$x_t = (1-B)^{-d} a_t. \tag{2.4}$$

We will discuss both Eqs. (2.3) and (2.4) later. First we start with equation (2.3)

$$(1-B)^d x_t = a_t.$$

We can represent fractionally differenced white noise (FDWN) as an infinite order autoregressive process

$$x_t = \sum_{k=0}^{\infty} \pi_k x_{t-k} \; + \; a_t.$$

Using the binomial expansion we can write

$$(1-B)^d = \{1 - dB + \frac{d(d-1)B^2}{2!} - \frac{d(d-1)(d-2)B^3}{3!} + ...\}. \tag{2.5}$$

Equation(2.5) can also be represented in terms of Hypergeometric function

$$(1-B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(k-d)B^k}{\Gamma(k+1)\Gamma(-d)}. \tag{2.6}$$

Using equation(2.6) we can calculate the value of π_k

$$\pi_k = \frac{\Gamma(k-d)}{\Gamma(-d)\Gamma(k+1)}. \tag{2.7}$$

Similarly we can also represent FDWN as an infinite order moving average process (2.4).

$$\begin{aligned} x_t &= (1-B)^{-d} a_t \\ &= \sum_{k=0}^{\infty} \psi_k a_{t-k} \end{aligned} \tag{2.8}$$

$$(1-B)^{-d} = \{1 + dB + \frac{d(d-1)B^2}{2!} + \frac{d(d-1)(d-2)B^3}{3!} + \dots\dots\dots\dots\dots\} \tag{2.9}$$

$$(1-B)^{-d} = \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)} B^k \tag{2.10}$$

Using above equation we can calculate the value of ψ_k

$$\psi_k = \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)}. \tag{2.11}$$

Based on the results derived by Granger(1980) the autocorrelation function of Fractional white noise is

$$\rho_k = \frac{\Gamma(k+d)\Gamma(1-d)}{\Gamma(k-d+1)\Gamma(d)}. \tag{2.12}$$

For large value of k, this can be approximated as

$$\rho_k = \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1}. \tag{2.13}$$

As we have already mentioned the main characteristic of *Long memory process* is that autocorrelations ρ_k at very long lags are nonzero and hence

$$\sum_{k=-\infty}^{\infty} |\rho_k| = \infty \tag{2.14}$$

Now using equation(2.13) and (2.14) we obtain

$$-1 < 2d - 1 < 0 \Longrightarrow d \in (0, 0.5).$$

In this section we have seen the formal definition of long memory time series, mathematical model to represent the long memory time series and their properties.

Chapter 3

Return interval distribution of long memory process with two scaling exponents

We consider long memory time series with two different scaling exponents. Detrended fluctuation analysis [5][4] is one of the most widely used techniques to identify the presence of long memory in a given time series. In Figure 3.1, we show the fluctuation function. It displays two different slopes α_1 and α_2. These slopes are known as DFA exponent. DFA exponent α is related to autocorrelation exponent γ (used in equation(2.1)) according to the relation given in equation (3.1).[9]

$$\gamma = 2 - 2\alpha \tag{3.1}$$

In Figure(3.1), the $x-$axis represents time scale n and $y-$axis $F(n)$ represents averaged fluctuation function. We can also see a crossover location where scaling exponent changes. Our probability model is the statment that

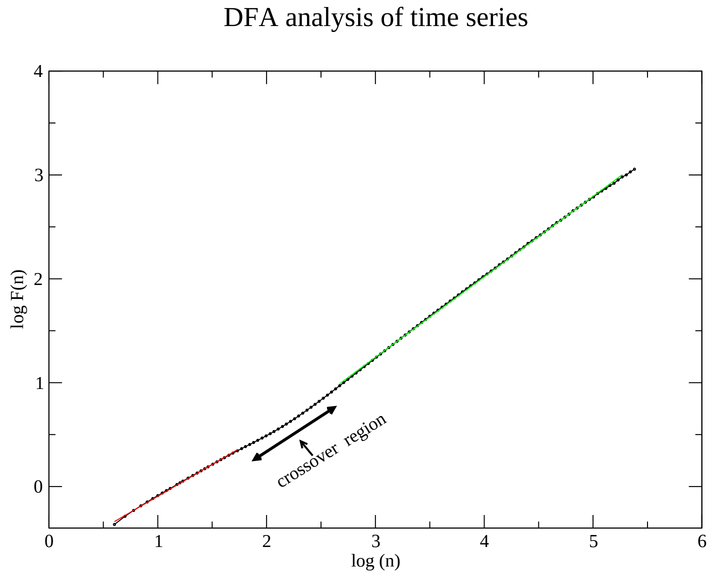


Figure 3.1: *DFA figure for time series with two scaling exponent*

for a stationary gaussian process with long memory, given an extreme evant at time $t = 0$, the probability to find an extreme event at $t = r$ is given by

$$P_{ex}(r) = \begin{cases} a_1 r^{-(2\alpha_1 - 1)} = a_1 r^{-(1-\gamma_1)} & \text{for } 0 < r < n_x \\ a_2 r^{-(2\alpha 2 - 1)} = a_2 r^{-(1-\gamma 2)} & \text{for } n_x < r < \infty \end{cases} \tag{3.2}$$

where $0.5 < \alpha_1, \alpha_2 < 1$ are DFA exponents and $0 < \gamma_1, \gamma_2 < 1$, a_1, a_2 are normalization constant which we will fix later. According to Equation(3.2) after an extreme event it is highly probable to expect the next event to be an extreme one too; and this is a resonable result for long memory persistent time series. After every extreme event one resets the time to zero and process starts again, independent of previous return interval(actually it is an assumption as we will further see during our numerical simulations that even return intervals have long memory).The range of γ is such that after a finite return intervals the process will stop. Next we will calculate, given an extreme event at $t = 0$, what will be the probability such that there is no extreme event in the interval $(0, r)$. To calculate this, we will first divide the interval r into m subintervals indexed by $j = 0, 1, 2, 3, ...(m-1)$ and then we calculate the probability in each subinterval. We have two cases, $case(1)$ is that when $r \in (0, n_x)$ and in $case(2)$ $r \in (n_x, \infty)$

For $case(1)$: $0 < r < n_x$

Here for the j^{th} subinterval,the probability of extreme event is given by (using Trapezoidal rule for integration[12])

$$h_1(j) = \frac{a_1 r}{m}(\frac{(j+1)r}{m})^{-(1-\gamma_1)} + \frac{a_1 r}{2m}[(\frac{jr}{m})^{-(1-\gamma_1)} - (\frac{(j+1)r}{m})^{-(1-\gamma_1)}] \quad (3.3)$$

after simplying this expression,the probability that no extreme event occurs in the j^{th} subinterval is given by

$$1 - h_1(j) = 1 - \frac{a_1 r}{2m}(\frac{r}{m})^{-(1-\gamma_1)}[(j+1)^{-(1-\gamma_1)} + j^{-(1-\gamma_1)}] \quad (3.4)$$

so probability of no extreme event in $(0, r)$

$$P_{noex}(r) = \lim_{m \to \infty} \prod_{j=0}^{m-1} (1 - h_1(j)) \quad (3.5)$$

we require the probability $P(r)dr$ that given an extreme event at $t = 0$, no extreme event occurs in $(0, r)$ and an extreme event occur in the infinitismal interval $r + dr$. This will be simply the product of $P_{noex}(r)$ with the probability $P_{ex}(r)$. This can be wirtten as

$$P(r)dr = P_{noex}(r)P_{ex}(r)dr \quad (3.6)$$

$$= \lim_{m \to \infty} [1 - \phi_1][1 - \phi_1(2^{-(1-\gamma_1)} + 1)]$$

$$\times[1-\phi_1(3^{-(1-\gamma_1)}+2^{-(1-\gamma_1)})]\dots\dots\dots$$

$$\times[1-\phi_1(m^{-(1-\gamma_1)}+(m-1)^{-(1-\gamma_1)})]a_1 r^{-(1-\gamma_1)}dr \qquad (3.7)$$

$$where \quad \phi_1 = \frac{a_1}{2}[\frac{r}{m}]^{\gamma_1}$$

The value of m can be extremely large and equation(3.7) can be written in a simplified form as

$$P(r)dr = \lim_{m\to\infty} exp[-\frac{a_1}{2}(\frac{r}{m})^{\gamma_1}\{2H_{m-1}^{\gamma_1-1}+m^{-(1-\gamma_1)}\}]a_1 r^{-(1-\gamma_1)}dr \qquad (3.8)$$

Where $H_{m-1}^{(\gamma_1-1)}$ is the generalized Harmonic number[11]. When we take the limit $m\to\infty$, we get

$$\lim_{m\to\infty}\frac{H_{m-1}^{\gamma_1-1}}{m^{\gamma_1}} = \frac{1}{\gamma_1}, \quad 0<\gamma_1<1. \qquad (3.9)$$

Using equation (3.8) and (3.9) we obtain the following results for return interval distribution

$$P(r)dr = a_1 r^{-(1-\gamma_1)}e^{-(a_1/\gamma_1)r^{\gamma_1}}dr \qquad (3.10)$$

For $case(2) : n_x < r < \infty$

Now in case(2) we will again follow the same kind of procedure that we have followed earlier to solve our problem in case (1) but the difference will start coming into the picture from equation (3.5) onwards. As we recall equation(3.5) is:

$$P_{noex}(r) = \lim_{m\to\infty}\prod_{j=0}^{m-1}(1-h_1(j)).$$

But now r is of length greater than n_x. Hence in place of equation(3.5) we have to use a new equation which reflects the problem that we are trying to solve in case(2). Now, the correct equation will be

$$P_{noex}(r) = \lim_{m\to\infty}[\prod_{j=0}^{\lambda}(1-h_1(j))\prod_{j=\lambda+1}^{m-1}(1-h_2(j))] \tag{3.11}$$

where,

$$1-h_1(j) = 1-\frac{a_1 r}{2m}(\frac{r}{m})^{-(1-\gamma_1)}[(j+1)^{-(1-\gamma_1)}+j^{-(1-\gamma_1)}]$$

and

$$1-h_2(j) = 1-\frac{a_2 r}{2m}(\frac{r}{m})^{-(1-\gamma_2)}[(j+1)^{-(1-\gamma_2)}+j^{-(1-\gamma_2)}]$$

We require the probability $P(r)dr$ that given an extreme event at $t=0$,no extreme event occurs in $(0,r)$ and an extreme event occur in the infinitismal interval $r+dr$. this will be simply the product of $P_{noex}(r)$ with the probability $P_{ex}(r)$. This can be written as

$$P(r)dr = P_{noex}(r)P_{ex}(r)dr \tag{3.12}$$

now use equation(3.11) and equation(3.2) we can write equation(3.12) as

$$\begin{aligned}P(r)dr = \lim_{m\to\infty}&[(1-h_1(0))(1-h_1(1))....(1-h_1(\lambda))\\ &\times(1-h_2(\lambda+1))(1-h_2(\lambda+2))....(1-h_2(m-1))]\\ &\times(a_2 r^{-(1-\gamma_2)})dr\end{aligned} \tag{3.13}$$

equation(3.13) can also be written as:

$$P(r)dr = \lim_{m\to\infty}[[(1-h_1(0))(1-h_1(1)).....(1-h_1(\lambda))]$$

$$\times \frac{[(1-h_2(0))(1-h_2(1))....(1-h_2(\lambda))....(1-h_2(m-1))]}{[(1-h_2(0))(1-h_2(1))....(1-h_2(\lambda))]}$$
$$\times a_2 r^{-(1-\gamma_2)} dr$$

which is equivalent to writing:

$$P(r)dr = C \lim_{m\to\infty} [\prod_{j=0}^{m-1} (1-h_2(j))] a_2 r^{-(1-\gamma_2)} dr \tag{3.14}$$

where C is:

$$C = \lim_{m\to\infty} \frac{[(1-h_1(0))(1-h_1(1))....(1-h_1(\lambda))]}{[(1-h_2(0))(1-h_2(1))....(1-h_2(\lambda))]} \tag{3.15}$$

repeating the same kind of calculations that we have done for case(1) we can write:

$$P(r)dr = C \lim_{m\to\infty} exp\Big[-\frac{a_2}{2}(\frac{r}{m})^{\gamma_2}\{2H_{m-1}^{\gamma_2-1} + m^{-(1-\gamma_2)}\}\Big] a_2 r^{-(1-\gamma_2)} dr \tag{3.16}$$

again using equation(3.9) we can show that:

$$P(r)dr = C a_2 r^{-(1-\gamma_2)} e^{-(a_2/\gamma_2) r^{\gamma_2}} dr \tag{3.17}$$

So final results for return interval distribution for extreme events are

$$P(r) = \begin{cases} (a_1 r^{-(1-\gamma_1)} e^{-(a_1/\gamma_1) r^{\gamma_1}}) & \text{for } 0 < r < n_x \\ C a_2 r^{-(1-\gamma_2)} e^{-(a_2/\gamma_2) r^{\gamma_2}} & \text{for } n_x < r < \infty \end{cases} \tag{3.18}$$

In equation(3.18) there are three unknowns a_1,a_2 and C. To solve for these values we can use three equations. First of this is the normalization equation:

$$\int_0^\infty P(r)dr = 1. \tag{3.19}$$

We will get second equation by normalizing $\langle r \rangle$ to unity

$$\int_0^\infty rP(r)dr = 1 \tag{3.20}$$

and the third equation will be obtained using continuity condition for equation(3.2)

$$\begin{cases} a_1 r^{-(1-\gamma_1)} = a_2 r^{-(1-\gamma_2)} \qquad \text{at } r = n_x \\ a_1 n_x^{-(1-\gamma_1)} = a_2 n_x^{-(1-\gamma_2)} \end{cases} \tag{3.21}$$

now using all three equations given above, we can find values of a_1,a_2 and C in terms of γ_1,γ_2 and n_x

using equation(3.19):

$$\int_0^\infty P(r)dr = \int_0^{n_x} a_1 r^{-(1-\gamma_1)} e^{-(a_1/\gamma_1)r^{\gamma_1}} dr + \int_{n_x}^\infty C a_2 r^{-(1-\gamma_2)} e^{-(a_2/\gamma_2)r^{\gamma_2}} dr = 1$$

This integral can be done by substituting $t = e^{-(a_1/\gamma_1)r^{\gamma_1}}$. This leads to a simple equation

$$Ce^{-(a_2/\gamma_2)n_x^{\gamma_2}} = e^{-(a_1/\gamma_1)n_x^{\gamma_1}}. \tag{3.22}$$

using equation(3.20):

$$\int_0^\infty rP(r)dr = \int_0^{n_x} ra_1 r^{-(1-\gamma_1)} e^{-(a_1/\gamma_1)r^{\gamma_1}} dr + C\int_{n_x}^\infty ra_2 r^{-(1-\gamma_2)} e^{-(a_2/\gamma_2)r^{\gamma_2}} dr = 1$$

to solve the above integral we should use integration by parts, after a very lengthy solving we will end up with not so pretty equation given below:

$$C(\gamma_2/a_2)^{1/\gamma_2} \frac{n_x E_{\frac{\gamma_2-1}{\gamma_2}}(n_x^{\gamma_2})}{\gamma_2} - (\gamma_1/a_1)^{1/\gamma_1} \frac{n_x E_{\frac{\gamma_1-1}{\gamma_1}}(n_x^{\gamma_1})}{\gamma_1} = 1 \tag{3.23}$$

Here $E_n(x) = \int_1^\infty \frac{e^{-xt}}{t^n} dt = \int_0^1 e^{-x/\eta} \eta^{(n-2)} d\eta$

$E_n(x)$ is known as exponential integral function.

Now using equation(3.21), (3.22) and (3.23), we can solve for values of a_1 ,a_2 and C. Since there are three variables and three equations we can get analytical solutions, but the form of equations is very complex. Hence, we prefer to solve them numerically. Hence with the known values of γ_1, γ_2 and n_x, we can obtain the corresponding values of a_1,a_2 and C.

Chapter 4

Simulation results

In previous chapter, we have solved our problem of return interval distribution analytically. To support those analytical results we would like to mention a model which will artificially generate a time series with different scaling exponents. The main idea behind this model is based on the discussions presented in in chapter(2)

Before describing our model we would like to mention that there are variety of models proposed to generate time series with two different exponents [10]. Quite similar to these other models, we also use use fractional differencing concept.

What we needed in our project work was the return interval distribution of time series, so our main focus is on return intervals, not on the time series. So we compromise just a bit with the model to generate time series within our specific needs.

Model described by Boris Podobnik[10] was taking almost sixteen hours to complete one simulation and if we want to generate an ensemble of hundred time series, it will take a really long time and hence it will be computationally very expensive. Hence we will use a different model which takes significantly less time and also generates time series with two different scaling exponents. Next we will discuss the model that we have used to generate the time series.

Step 1:

set the length of time series, say, $l = 10^5$.

Step 2:

generate a series of random numbers y_i $i = 0......(l-1)$ which follow gaussian distribution with mean 0 and variance 1

Step 3:

generate a series of coefficients defined as:

$$C_i^\alpha = \frac{\Gamma(i-\alpha)}{\Gamma(-\alpha)\Gamma(i+1)} = -\frac{\alpha}{\Gamma(1-\alpha)}\frac{\Gamma(i-\alpha)}{\Gamma(i+1)}$$

$$\alpha = \begin{cases} \alpha_1 & \text{for } 0 < r < n_x \\ \alpha_2 & \text{for } n_x < r < \infty \end{cases} \tag{4.1}$$

Both α_1 and α_2 belong to the interval $(-0.5, 0)$

The asymptotic behaviour of C_i^α for large i can be written as

$$C_i^\alpha \simeq -\frac{\alpha}{\Gamma(1-\alpha)} i^{-(1+\alpha)} \quad for \quad i \gg 1$$

Step 4:

Now, get a series y_i^α using y_i and C_i^α according to the relation

$$y_i^\alpha = \sum_{j=0}^{i} y_{i-j} C_j^\alpha \qquad i = 0.....(l-1) \tag{4.2}$$

When we do DFA on the series y_i^α, we will obtain the figure given below which shows that the generated time series has two DFA exponent (two different scaling exponent). This was the desired characteristic for the time series that we want to generate. So for the model described above according to equation(4.2) if we put $n_x = 30$ and use two different values of α let say α_1 (for $0 \leq i \leq n_x$) and α_2 (for $n_x \leq i \leq (l-1)$), we can generate a time series with two different scaling exponents. After getting the time series we will apply DFA analysis on the time series. If we look at the DFA figure given below we will find that there are two different slopes and the crossover location of slope is consistent with the crossover location that we have supplied to generate the time series (crossover location is $2n_x$)[10] Now

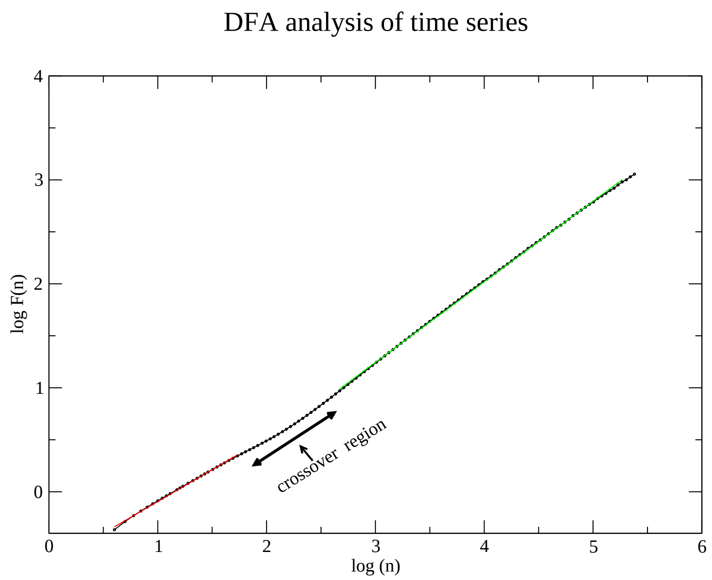


Figure 4.1: *DFA of time series with* $n_x = 30$,$\alpha_1 = 0.61$ *and* $\alpha_2 = 0.78$

since we have the time series in hand and we also know the autocorrelation exponents $\gamma_1 = 0.78$ and $\gamma_2 = 0.44$(use relation $\gamma = 2 - 2\alpha$,here α is the DFA exponent...also cite the paper for this relation), we can move on to calculate the return interval distribution of extreme events for this time series. In chapter(1) we have already described how we are going to calculate the return interval distribution for the time series. We will use the same approach here as well. If we notice the Figure(4.2) we see that there is a breakpoint in

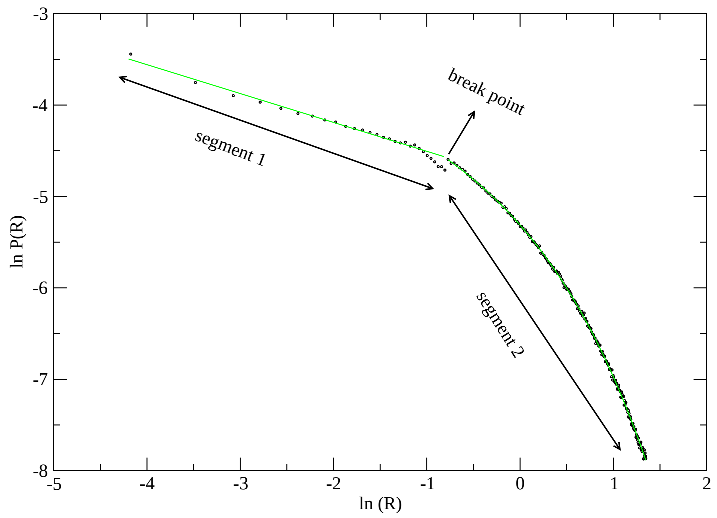


Figure 4.2: *return interval distribution of time series with* $n_x = 30$ *and* $q = 2.2$ *and* R *is the scaled return interval defined as* $r/\langle r \rangle$. *Points represent numerical data, line represents fitting of Equation(4.3)*

the figure at $R = n_x/\langle r \rangle$ (for the specfic example shown above the value of $\langle r \rangle$ is 65 units) So we can see that there are two segments in return interval distribution figure; first segment contains all those return intervals for which $r < n_x$ and the second segment contains the return intervals for which $r > n_x$. The return interval distribution figure that we have shown above is the result of our analysis on an ensemble of hundread time series. This is done to avoid the excessive fluctuation for the return interval distribution calculation. Next we have tried to fit both the segments of return interval distribution with a curve of the form,

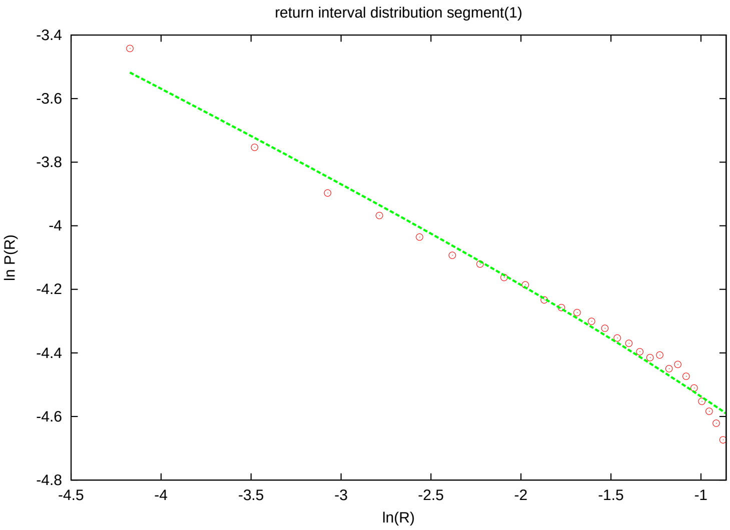
$$P(r) = ar^{-(1-\gamma)}e^{-(c/\gamma)r^{\gamma}}. \tag{4.3}$$

These fits are shown in Figure(4.3(a)) and Figure(4.3(b)). One can raise a quesition about equation(4.3) that why should two different values a and c be used, why not one single value a or c ? The explanation for the question is that since return interval distribution also follow long memory and we have taken the assumption that return intervals are completely uncorrelated. Hence, a significant difference will arise between our numerical and analytical results. Considering the long memory dependence of return intervals will be a problem in itself and we are not going to solve that issue in this report. So in our results, we assume that return intervals are completely uncorrelated. One more problem with the equation(4.3) is that the γ values are not completely consistent with the γ values based on the DFA exponents of the time series. There are two possible cause for this inconsistency:

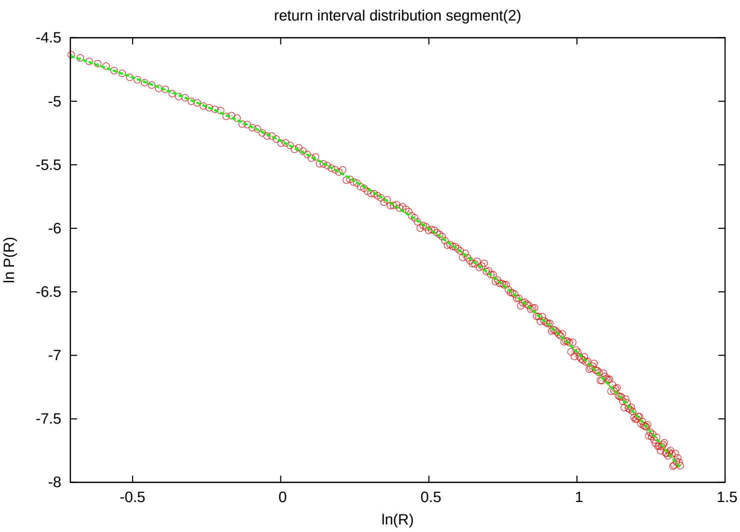
(1) There is a transition region for the crossover, ideally the transition should be sharp enough to get better results. We can't control the adverse effect of this problem on our results.

(2)We are not able to understand what is the effect of different threshold values(q) on the curve described by equation(4.3).

So the main conclusion that we can derive from this section is that both the return interval segments follow a distribution which is of the form equation(4.3). We can say that return interval distributions still follow a form which contains product of power law and a stretched exponential. These results are kind of same as results shown in Santhanam et,al(cite his reference here), the only difference is the apperance of a break point which is consistent with



(a) segment 1



(b) segment 2

Figure 4.3: *Return interval distribution of long memory time series with two scaling exponent. Circle represents numerical simulations,broken line represents analytical results*

the crossover in DFA figure. In next section we will generate a long memory probability process with two scaling exponents. This closely follows the theoretical assumptions and hence we can expect a better agreement with the analytical results.

Chapter 5

Long memory probability process with two scaling exponents

In previous section, we have seen that our numerical results are deviate to some extent from our analytical results. One of the reason that we have suggested for this discrepancy is the presence of long memory in return intervals. So if we somehow manage to get rid of this problem our analytical results should be consistent with the numerical results. Now we will try to numerically simulate the probability process in equation(3.2), we first determine the constants a_1 and a_2 by normalizing it in the region $k_{min} = 1$ and k_{max}. Using equation(3.2) we can set the normalization condition,

$$\int_1^{k_{max}} P_{ex}(r)dr = \int_1^{n_x} a_1 r^{-(1-\gamma_1)} dr + \int_{n_x}^{k_{max}} a_2 r^{-(1-\gamma_2)} dr = 1 \qquad (5.1)$$

except this normalization condition we will also use the continuity condition 3.21 to calculate the values of a_1 and a_2.

After solving equation(3.21) and (5.1) the values of a_1 and a_2 are given below.

$$a_1 = \frac{1}{[\frac{n_x^{\gamma_1}}{\gamma_1} - \frac{1}{\gamma_1} + \frac{k_{max}^{\gamma_2} n_x^{\gamma_1-\gamma_2}}{\gamma_2} - \frac{n_x^{\gamma_1}}{\gamma_2}]} \tag{5.2}$$

$$a_2 = \frac{1}{[\frac{n_x^{\gamma_2}}{\gamma_1} - \frac{n_x^{\gamma_2-\gamma_1}}{\gamma_1} + \frac{k_{max}^{\gamma_2}}{\gamma_2} - \frac{n_x^{\gamma_2}}{\gamma_2}]} \tag{5.3}$$

So now we have the values of a_1 and a_2,we can use the probability distribution given below.

$$P_{ex}(r) = \begin{cases} a_1 r^{-(2\alpha_1-1)} = a_1 r^{-(1-\gamma_1)} & \text{for } 1 < r < n_x \\ a_2 r^{-(2\alpha 2-1)} = a_2 r^{-(1-\gamma 2)} & \text{for } n_x < r < k_{max} \end{cases} \tag{5.4}$$

where $r = 1, 2, 3......$ Now generate a random number ξ_r from a uniform distribution at every r and compare it with the value of $P(r)$. A random number is accepted as an extreme event if $\xi_r < P(r)$ at any given value of r. If $\xi_r \geq P(r)$, then it is not an extreme event. Using this procedure we can generate a series of extreme events which follow equation(5.4). Now we will calculate the return interval distribution after scaling it by average return interval.

In the Figure(5.1) we have two segments for return interval distribution. According to the result shown in equation(3.18), we can fit both of these segments. In equation(3.18) we have three constants a_1, a_2 and C, we have already solved for the values of a_1 and a_2 in equation(5.2) and equation(5.3). To calculate the value of C, we will use the normalization condition for total probability,

$$\begin{aligned} \int_1^{k_{max}} P(r)dr &= \int_1^{n_x} P(r)dr + \int_{n_x}^{k_{max}} P(r)dr = 1 \\ &= \int_1^{n_x} a_1 r^{-(1-\gamma_1)} e^{-(a_1/\gamma_1) r^{\gamma_1}} dr + \int_{n_x}^{k_{max}} C a_2 r^{-(1-\gamma_2)} e^{-(a_2/\gamma_2) r^{\gamma_2}} \end{aligned}$$

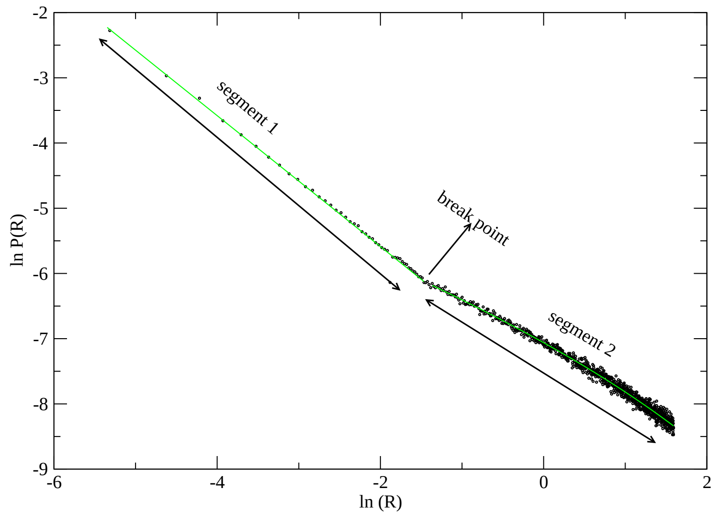
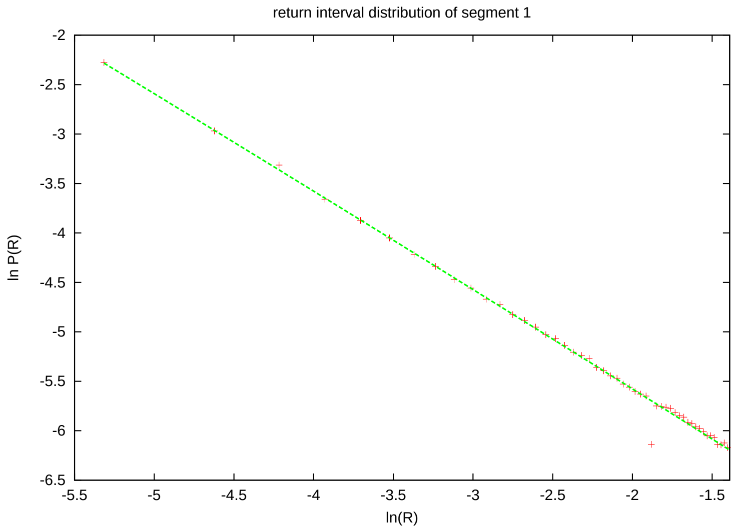


Figure 5.1: *return interval distribution of long memory probability process for* $\gamma_1 = 0.1$,$\gamma_2 = 0.5$,$n_x = 50$ *and* $k_{max} = 1000$*, points represent numerical resuts, line represents analytical results given in Equation(3.18)*

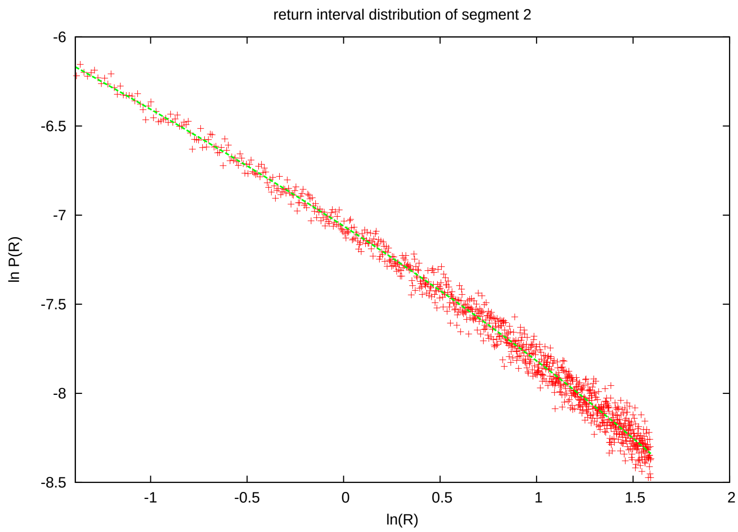
On solving the integrals given above we will get an equation containing a_1, a_2 and C.

$$[e^{-(a_1/\gamma_1)} - e^{-(a_1/\gamma_1)n_x^{\gamma_1}}] + C[e^{-(a_2/\gamma_2)n_x^{\gamma_2}} - e^{-(a_2/\gamma_2)k_{max}^{\gamma_2}}] = 1 \tag{5.5}$$

If we use equation(5.2), (5.3) and substitute the values of a_1 and a_2 in equation(5.5), we will get C in terms of n_x, γ_1 and γ_2. Since the equations describing C, a_1 and a_2 are of very complex form, it will be better to solve them numerically using the values of γ_1, γ_2 and n_x. As we have seen in the Figure(5.2(a)) and Figure(5.2(b)), we are able to fit both of these segments according to results shown in equation(3.18). But if we inspect these figures closely we will find that for first segment the the actual fitting is not $a_1 r^{-(1-\gamma_1)} e^{-(a_1/\gamma_1)r^{\gamma_1}}$, instead the fiting is something like $ar^{-(1-\gamma_1)} e^{-(b/\gamma_1)r^{\gamma_1}}$.



(a) segment 1 for $\gamma_1 = 0.1$



(b) segment 2 for $\gamma_2 = 0.5$

Figure 5.2: *return interval distribution of long memory probability process (+ represents numerical results, broken line represents analytical results)*

So for this probability process we need to explain,why the return interval distribution expression of segment(1) have two different constants a and b(inplace of one single constant a_1 according to equation(3.18)). According to equation(5.4), the minimum size of return interval possible is 1 unit. If we do scaling on this by the average return interval $\langle r \rangle$ then the scaled minimum return interval will be $1/\langle r \rangle$. So in equation(3.19) and (3.20), we should replace the lower limit of integral by $1/\langle r \rangle$ in place of 0. It also reflects the general idea that all power laws in practice have a lower bound. So the replacement of lower limit in the integrals will lead to different constants a and b instead of a single constant a_1.

Chapter 6

Conclusion

We have studied the distribution of return intervals for long memory process with two different scaling exponents. We have obtained an analytical expression for the return interval distribution and verified it with simulations. We have shown that for a long memory time series with two different scaling exponent there will be a crossover point in the return interval distribution. We have shown that if a time series has different scaling exponents, then this is also reflected in the return interval distribution. For each scaling exponent there will be a corresponding segment in the return interval distribution and the common thing about all the segments is that all of them still follow a distribution which is the product of a power law and stretched exponential. The only difference is the scaling exponent that is going to appear for different segments. The previous studies in the field of extreme events mainly focus on time series with single scaling exponent although many of the time series observed in nature contain more than one scaling exponent. So the results that we have shown in this report could help us deal with real life series more accurately. Although the results shown in this report are quite encouraging, there are a many issues which are needed to be resolved for much better analysis; (a) the model that we have used to generate time series with more than one scaling exponent need a fine tuning so that we can test our analytical results more accurately, (b) we should also think of the effects of long memory in return intervals itself. Future research in this field should address some of these issues.

Bibliography

[1] Rosario N. Mantegna, H. Eugene Stanley. *An introduction to econophysics:correlations and complexity in finance.* Cambridge university press

[2] G. Rangarajan, M. Ding. *Processes with long range correlations:Theory and application* Springer

[3] Jonathan D. Cryer, Kung-Sik Chan. *Time series analysis, with application in R.* Springer

[4] Peng C-K, Havlin S, Stanley HE, Goldberger AL. *Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series.* Chaos 1995;5:82-87.

[5] Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. Phys Rev E 1994;49:1685-1689.

[6] Roman H. E and Porto M, *Fractional derivatives of random walks:Time series with long time memory* arXiv:0806.3171v1

[7] Richard T. Baillie,1996.*Long memory process and fractional integration in econometrics.* Journal of econometrics 73 (1996) 5 59

[8] M. S Santhanam, Holger Kantz. *Return interval distribution of extreme events and long-term memory.* Phys Rev E 78, 051113 (2008)

[9] Govindan Rangarajan, Mingzhou Ding. *Integrated approach to the assessment of long range correlation in time series data* Phys Rev E 61, 5, 4991-5001, (2000)

[10] Boris Podobnik, Ivo Grosse, H. Eugene Stanley. *Stochastic processes with power-law stability and a crossover in power-law correlations* Physica A 316 (2002) 153 – 159

[11] Ronald L. Graham, Donald E. Knuth, Oren Patashnik. *Concrete Mathematics* Addison-Wesley publishing company

[12] William H. press, Saul A. Teukolsky, William T. Veterling, Brian P. Flannery. *Numerical recipes in c++: The art of scientific computing.* Cambridge university press