CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

FACULTY
OF NUCLEAR SCIENCES
AND PHYSICAL
ENGINEERING

DOCTORAL
THESIS

2022

ALTERNATIVE APPROACHES OF RENYI DIMENSION ESTIMATION

MARTIN
DLASK

Bibliografický záznam

Autor	Ing. Martin Dlask, České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, katedra softwarového inženýrství
Název práce	Alternativní přístupy k odhadu Rényiho dimenze
Studijní program	Aplikace přírodních věd
Studijní obor	Matematické inženýrství
Školitel	doc. Ing. Jaromír Kukal, PhD., České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, katedra softwarového inženýrství
Školitel specialista	prof. Ing. Pavel Sovka, CSc., České vysoké učení technické v Praze, Fakulta elektrotechnická, katedra teorie obvodů
Akademický rok	2021/2022
Počet stran	176
Klíčová slova	bodová množina, zlomkový Brownův pohyb, Rényiho dimenze, Hurstův exponent

Bibliographic entry

Author	Ing. Martin Dlask, Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Software Engineering
Title of Dissertation	Alternative Approaches of Renyi dimension estimation
Degree programme	Application of natural sciences
Field of Study	Mathematical Engineering
Supervisor	doc. Ing. Jaromir Kukal, PhD., Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Software Engineering
Supervisor specialist	prof. Ing. Pavel Sovka, CSc.,Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory
Academic Year	2021/2022
Number of Pages	176
Keywords	point set, fractional Brownian motion, Renyi dimension, Hurst exponent

Abstrakt

Práce se zabývá vývojem alternativních přístupů k odhadu Rényiho dimenze a Hurstova exponentu. Cílem je vypracovat metody, které by vedly k odstranění systematické chyby odhadu, snížení standardní odchylky odhadu a zobecnění přístupů k odhadu Rényiho dimenze pro bodové množiny a zlomkové procesy. Práce představuje novou metodu rotačního spektra pro odhad korelační dimenze a metodu modifikované Rényiho entropie pro odhad Rényiho dimenze. Jejich vlastnosti jsou studovány teoreticky i experimentálně. V oblasti odhadu Hurstova exponentu jsou vyvinuty metody modifikovaného průchodu nulou, zlomkového Brownova mostu a autokorelační přístup s využitím Wishartova rozdělení. Všechny metody jsou podrobeny statistickému testování na množinách se známou dimenzí s cílem ověřit jejich vydatnost, nestrannost a konzistenci. Metody odhadu dimenze jsou aplikovány na reálná data pro detekci Alzheimerovy choroby ze signálu EEG, měření predikovatelnosti indexů akciových trhů, studium signálu autolumunescence a detekci rakoviny prsu z mamografie.

Abstract

The work presents alternative approaches to Renyi dimension and Hurst Exponent estimation. The aim is to develop methods leading to bias removal, reduction of the standard deviation of the estimate and generalization of the approaches to Renyi dimension estimation both for point sets as well as fractional processes. The novel method of rotational spectrum for correlation dimension estimation and the modified Renyi entropy method for general Renyi dimension estimations are developed and their properties are theoretically and experimentally verified. For analysis of fractional processes, the method of revisited zero-crossing, fractional Brownian bridge and autocorrelation using Wishart distribution are presented. All the methods are statistically tested on artificial sets with known dimension and their unbiasedness, efficiency and consistency is verified. The methods were applied to real data and proven to be suitable for Alzheimer disease detection from EEG signal, evaluating predictability of stock market indices, analysis of autoluminiscence signal and breast cancer detection from mammography.

Acknowledgements

Hereby I would like to thank and show my gratitude to Jaromir Kukal for receiving a great deal of support over the years, for willingness to consult regularly, for a large number of stimulating ideas and valuable discussions over interesting topics. It was truly a privilege to be mentored by someone who has such an extensive experience in the field of statistics and beyond.

Contents

List of Commonly Used Symbols

$\mathcal{B}(\vec{x}, \epsilon)$	ball with radius ϵ centered at $\vec{x}$
$B_H(t)$	fractional Brownian motion
$\mathbb{C}$	set of complex numbers
$C(r)$	correlation sum
card	cardinality of a point set
diam	diameter of a point set
D_0	capacity dimension
D_1	information dimension
D_2	correlation dimension
D_α	Renyi α-dimension
D_α^*	modified Renyi α-dimension
D_H	Hausdorff dimension
D_T	topological dimension
$\mathcal{DFT}$	discrete Fourier transform
$\mathcal{DFT}^{-1}$	inverse discrete Fourier transform
E	expected value
$f(z\|p)$	probability density function with parameter p
$\mathcal{F}$	point set
$\mathcal{F}_n$	n-dimensional Fourier transform
$\mathcal{F}_n^{-1}$	n-dimensional inverse Fourier transform
$\mathcal{G}_\epsilon$	grid of size ϵ
$G_H(t)$	fractional Brownian noise
H	Hurst exponent
H_0	null hypothesis
H_0	Hartley entropy
H_1	Shannon entropy
H_2	Collision entropy
H_∞	minimum entropy
H_α	Renyi entropy
H_α^*	modified Renyi entropy
$\mathcal{H}_s(\mathcal{F})$	Hausdorff's s-dimensional measure of point set $\mathcal{F}$
I(...)	indicator function
ln	natural logarithm
log	decimal logarithm
$\log_2$	binary logarithm

$L_\alpha(t)$	Levy flight
$M(t)$	Brownian bridge
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of non-negative integers
$\mathrm{N}(0,1)$	standard normal distribution
$N_\delta(\mathcal{F})$	δ-coverage of set $\mathcal{F}$
$\mathcal{O}(\ldots)$	big O notation
p-val	p-value
$P(\vec{\omega})$	Power spectrum of a point set
$\mathrm{P}(\ldots)$	event probability
$\mathbb{R}$	set of real numbers
$\mathbb{S}$	scatter matrix
$S(\Omega)$	rotational spectrum of a point set
$\mathcal{S}_{n-1}$	n-dimensional sphere
$\mathrm{SO}(n)$	group of all rotations in $\mathbb{R}^n$ around origin
$T_H(t)$	fractional Brownian trace
$T(\vec{x})$	Takagi function
$\mathrm{tr}(\ldots)$	trace
$\mathrm{U}(\ldots)$	uniform distribution
var	variance
$W(\vec{x})$	Weierstrass function
$\mathbb{Z}$	set of integers
$(\mathcal{A}_i)$	sequence of sets
(a_i)	sequence of numbers
$(f \star g)$	convolution of functions f and g
$\delta_{i,j}$	Kronecker delta
ρ_k	autocorrelation coefficient
$\|...\|$	Euclidean norm
$\bigcup$	union
$\times$	Cartesian product
[...]	fractional part of a number

List of Abbreviations

AD	Alzheimer disease
BAL	biological autoluminiscence
BB	Brownian bridge
CI	confidence interval
CN	control normal
CR	confidence region
dfBB	differenced fractional Brownian bridge
EEG	electroencephalogram
fBm	fractional Brownian motion
FDR	false discovery rate
fGn	fractional Gaussian noise
GCE	generalized circulant embedding
ifGn	incremental fractional Gaussian noise
LR	likelihood ratio
MLE	maximum likelihood estimation
MSE	mean square error
pdf	probability density function
PMT	photomultiplier tube
SR	space reconstruction
TCE	traditional circulant embedding

List of Figures

List of Tables

Introduction

Fractal sets are a special class of structures in Euclidean space, having non-integer dimension that exceeds their standard topological dimension. Many natural phenomena look like they are fractals, but with respect to its formal definition, real fractal sets exist only in theory. The infinite amount of detail is the pre-requisite of having non-integer dimension and in real world, every structure has finite number of elements - in case of scientific computing, we are limited by finite memory of a computer, in case of natural phenomena, the scale as the measure of detail ends with the size of a cell in biological sense or size of atom in the physical sense. Therefore, there is most likely no structure in the world that would be a true fractal.

Why it is useful to develop the theory of fractal dimension estimation, when there are no true fractals in reality? In fact, many real-world phenomena, from the area of economics, hydrology, biomedicine and many more are *samples* of fractal sets. While we are never able to find a structure with infinite amount of detail, we assume the real data are uniformly distributed samples from the theoretical fractal structure and we can use those to be modelled using a continuous model. Computational experiments from the past have shown even when using simple techniques for the fractal dimension estimation, the structures significantly differ in their dimension and can be suitable for fractal modelling.

The concept of non-integer dimension is not new, the beginning of the investigations of the fractal sets started already in the 60s of the previous century, pioneered by Hurst [1] and Mandelbrot [2]. Since then, the non-integer dimension was examined both in theoretical and experimental way. There were developed many fractal sets and functions with known dimension that can nowadays serve as as good examples for testing purposes.

Looking at the current state of the art, there are various definitions of fractal dimensions as well as conventional methods for its estimation. To the most common definitions belong Hausdorff dimension, self-similarity dimension and Renyi dimension. Nowadays, any approach leads to several challenges, such as

- uniform generation of points from deterministic and random fractals,
- storing very detailed fractal structures in computer memory,
- precise estimation of any type of fractal dimension.

The first challenge exists already from the origination of chaos theory. There are many ways how to generate points in $\mathbb{R}^n$ – using the self-similarity and associated contraction mappings, one can generate the points in certain recurrence depth. The other possibility of generation of fractal functions is the uniform sampling of the function domain and the calculation of dependent values. A little bit more trickier is the generation of points from multidimensional processes since one needs to use specialized methods that are compliant with the distribution of the multidimensional random variable. Most methods, regardless the type of fractal dimension, require uniform sampling of points from the original set.

Storing the fractal set for analysis is also not an easy task. Since the fractal sets have usually infinite level of detail, it is not possible to store them in computers memory point by point. When measuring the data from real word, such as financial or biomedical data, the detail of the measurement (or sampling rate) and ability of storing this detail in computer memory plays an important role in the quality of subsequent estimate.

The biggest challenge that is not resolved yet for many types of fractal structures is the precise and unbiased estimate of its dimension. The current methods that are widely used both in mathematics and engineering are heavily dependent on the type of investigated structure. For fixed structure of data, such as time series, set sampling or binary picture, the methods can provide relatively accurate estimate of the dimension.

The research work described in this thesis is directly challenging the methods that are considered "traditional" for fractal dimension estimation and providing alternative approaches that are aimed to:

- reduce the bias from the estimation,
- diminish variance of the estimate,
- generalize the approaches to Renyi dimension estimation,
- provide an estimate even when data sample is small,
- provide dimension estimate of real signals and images.

The traditional methods are simple to use, but they come with a cost of a bias that is present in the estimate, especially when following conventional grid methods. The uncertainty of the estimates is high and therefore the conventional algorithms are not so useful for being applied on real data. The problem is present as well when the size of the input data is small. The general approach of the research methodology for developing new methods for Renyi dimension estimation always obeys following procedure:

1. Introduce a new definition of the dimension and examine its properties.
2. Theoretically prove that the new definition is equivalent to the traditional definition.

3. Define an estimation algorithm using the new definition of the dimension.
4. Examine this algorithm on simulated data with known dimension and compare to the traditional method of estimation.
5. Apply the methods to samples from real processes and/or images.

The thesis follows following order; Chapters 1-3 provide the overview of the current state of art with respect to the dimension definition, description of fractals with known dimension as well as the sketch of the traditional methods that are used for Renyi dimension estimation. Chapter 4 summarizes the thesis aims and describes the research goal. Chapters 5-9 describe novel methods for fractal dimension estimation, with the focus on correlation and Renyi dimension in case of point sets and with focus on Hurst exponent estimation in the case of random processes and time series. Experimental part is divided between chapters 10 and 11, while we are first demonstrating the performance of the developed methods on the fractals with known dimension and afterwards showing case studies of the applications on real data.

Chapter 1

Point Set can be a Fractal Set

This chapter provides an introduction to the fractal sets and summarizes the well-known fractal dimension types and their properties.

1.1 Topological Dimension

Sets in n-dimensional Euclidean space can be characterized by their topological dimension. This dimension always exists and its definition is based on covering. The topological dimension is the most common type of dimension and represents the natural understanding of the space measure. As a simple example, the topological dimension of finite number of points in Euclidean space equals zero, while the topological dimension of a finite set of lines equals one.

For set $\mathcal{A} \subset \mathbb{R}^n$ we define [3] the δ-coverage of set $\mathcal{A}$ as a countable sequence of open sets $(\mathcal{A}_i)$, $i = 1, \ldots, +\infty$, satisfying

- $\operatorname{diam}(\mathcal{A}_i) \leqslant \delta$ for each i,
- $\mathcal{A} \subset \bigcup_i \mathcal{A}_i$.

For any δ-coverage of the investigated set, we can always find a refinement, i.e. the ϵ-coverage for $\epsilon < \delta$. The topological dimension of set $\mathcal{A} \subset \mathbb{R}^n$ is the smallest non-negative integer m such that for any coverage of set A, there exists a refinement satisfying that every point from set A is contained in the intersection at most $m+1$ sets from the refinement.

The topological dimension is often denoted as D_{T}. For any set $\mathcal{A} \subset \mathbb{R}^n$, D_{T} is always non-negative integer smaller or equal to n. The topological dimension is important in the definition of a fractal sets.

1.2 Hausdorff Dimension

Hausdorff measure and dimension are the traditional descriptors of a fractal set. The Hausdorff s-dimensional measure of any set $\mathcal{F} \subset \mathbb{R}^n$ is defined [3] as

$$\mathcal{H}_s(\mathcal{F}) = \lim_{\delta \to 0} \inf \left\{ \sum_{i=1}^{+\infty} (\mathrm{diam}(\mathcal{A}_i))^s : (\mathcal{A}_i) \text{ is } \delta\text{-coverage of set } \mathcal{F} \right\}. \tag{1.1}$$

As easy to prove, the limit from the equation (1.1) always exists [3]. The interesting property of Hausdorff s-dimensional measure is that for every set, for most $s > 0$, the measure will be having values either 0 or $+\infty$. In fact, for any $\mathcal{F} \subset \mathbb{R}^n$, there exists certain $s_0 > 0$ such as $\mathcal{H}_s(\mathcal{F}) = +\infty$ for any $s < s_0$ and $\mathcal{H}_s(\mathcal{F}) = 0$ for $s > s_0$. Since the value of s_0 is unique for any Euclidean set, the Hausdorff dimension is defined as

$$D_{\mathrm{H}} = \inf \{ s : \mathcal{H}_s(\mathcal{F}) = 0 \text{ and } s \geqslant 0 \}. \tag{1.2}$$

The Hausdorff dimension is a theoretical construct and provides the foundation of fractal sets. Having both Hausdorff and topological dimension in hand, any set $\mathcal{F} \subset \mathbb{R}^n$ that fulfils the condition of $D_{\mathrm{T}} < D_{\mathrm{H}}$ is called a fractal set. Moreover, according to Sznirelman theorem [4] the Hausdorff dimension is always greater or equal than the topological dimension, therefore for a set to be qualified as fractal, strong inequality must occur between these two dimension types. The advantage of Hausdoff dimension is its general existence for any set, however the deduction of the dimension from the definition is fairly difficult. Therefore, there are other types of dimension that allows to estimate its lower and upper bound using simpler algorithms.

1.3 Capacity Dimension

Capacity dimension [5, 3, 6] (sometimes called box-counting dimension) is another type of fractal dimension that is defined again via coverings, however the covering sets can have grid structure and the definition does not employ any measure.

For non-empty and bounded set $\mathcal{F} \subset \mathbb{R}^n$ and parameter $\delta > 0$ we define $N_\delta(\mathcal{F})$ as the minimum number of sets with diameter at most δ that can cover the set $\mathcal{F}$, i.e. the set $\mathcal{F}$ is the subset of the union of the covering sets. The capacity dimension is based on the definition of lower and upper case, when as the covering diameter δ goes to zero. The lower capacity dimension is defined as

$$\underline{\dim_{\mathrm{B}}}(\mathcal{F}) = \liminf_{\delta \to 0^+} \frac{\log N_\delta(\mathcal{F})}{-\log \delta} \tag{1.3}$$

and the upper capacity dimension is defined as

$$\overline{\dim_{\mathrm{B}}}(\mathcal{F}) = \limsup_{\delta \to 0^+} \frac{\log N_\delta(\mathcal{F})}{-\log \delta}. \tag{1.4}$$

In case that $\overline{\dim_{\mathrm{B}}}(\mathcal{F}) = \underline{\dim_{\mathrm{B}}}(\mathcal{F})$, we define the capacity dimension of set $\mathcal{F}$ as

$$D_0 = \lim_{\delta \to 0^+} \frac{\log N_\delta(\mathcal{F})}{-\log \delta}. \tag{1.5}$$

For $\delta \to 0^+$, both nominator and denominator will be increasing their value to $+\infty$. Only when the upper dimension equals the lower dimension, the capacity dimension exists. This condition is, however, fulfilled in the majority of self-similar sets. Moreover, the the capacity dimension D_0 is the upper bound of the Hausdorff dimension, i.e. $D_{\mathrm{H}} \leqslant D_0$. For any set that satisfies the open set condition [7], the two dimensions are equal.

To simplify the calculation of capacity dimension of Euclidean set $\mathcal{F} \subset \mathbb{R}^n$, the conditions in the coverage can be further relaxed [3]. The parameter N_δ can denote the minimum number of closed (hyper)balls of radius δ that cover $\mathcal{F}$ as well as by the (hyper)cubes of side δ that cover $\mathcal{F}$. The covering cubes can be arranged into a grid, which can simplify the experimental estimation of the capacity dimension similarly as in Fig. 1.1.

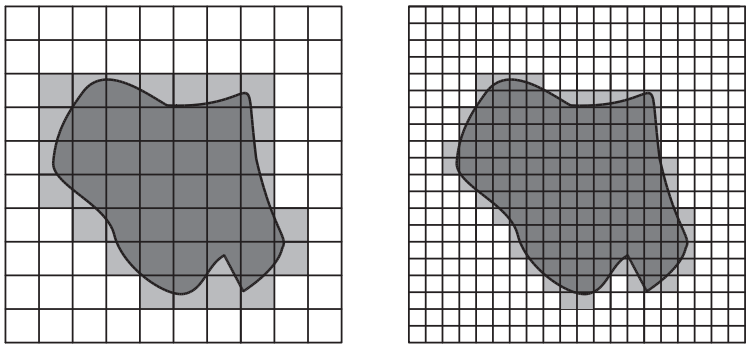


Figure 1.1: Set inside the grid of side $\delta = 1$ with $N_1 = 45$ (left) and $\delta = 1/2$ with $N_2 = 144$ (right).

1.4 Correlation Dimension

Correlation dimension is based on the calculation of Euclidean distance between points inside the investigated set $\mathcal{F} \subset \mathbb{R}^n$. To quantify the distances between the points, one can use the correlation sum $C(r)$ [8, 9] defined for any $r > 0$ as

$$C(r) = \lim_{N \to +\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I\left(\|\vec{x_i} - \vec{x_j}\| \leqslant r\right) \tag{1.6}$$

for $\vec{x_i}, \vec{x_j} \in \mathcal{F}$ and indicator function I.

Correlation sum indicates what is the probability that the distance between two points from the point set $\mathcal{F}$ is at most r. Therefore, this can be equivalently expressed as

$$C(r) = \underset{\vec{x},\vec{y}\in\mathcal{F}}{\mathrm{P}}(\|\vec{x}-\vec{y}\| \leqslant r)\,. \tag{1.7}$$

Using the correlation sum, one can define the correlation dimension as

$$D_2 = \lim_{r\to 0^+} \frac{\ln\, C(r)}{\ln\, r}, \tag{1.8}$$

if the limit exists. As easy to prove, for any set with finite number of points, the correlation dimension equals zero. The correlation dimension can serve as the lower bound for Hausdorff dimension estimation, e.g.

$$D_{\mathrm{T}} \leqslant D_2 \leqslant D_{\mathrm{H}} \leqslant D_0. \tag{1.9}$$

1.5 Renyi Dimension

The parameterized type of a fractal dimension is the Renyi dimension [10, 11] that generalizes the capacity and correlation dimensions. The calculation of the Renyi dimension is based on the Renyi entropy [11] H_α, which is a generalisation of the Shannon (H_1) [12], Hartley (H_0) [13] and collision [14] (H_2) entropies. The α-entropy is defined for $\alpha \geqslant 0$ as follows

$$H_\alpha(\vec{p}) = \frac{1}{1-\alpha}\ln\sum_{i=1}^{k} p_i^\alpha \tag{1.10}$$

for $\alpha \neq 1$ and

$$H_1 = \lim_{\alpha\to 1} H_\alpha = -\sum_{i=1}^{k} p_i \ln p_i \tag{1.11}$$

where k is the number of events and p_i are their probabilities satisfying $\sum p_i = 1$. The formulas (1.10) and (1.11) are frequently used in most sources, but they describe only a finite set of events with a possible extension to a countable case. A more general form which includes also an uncountable case is defined as

$$H_\alpha(\vec{p}) = \frac{1}{1-\alpha}\ln \mathrm{E}\ \left(p^{\alpha-1}\right) \tag{1.12}$$

for $\alpha \neq 1$ and

$$H_1 = \mathrm{E}\ (-\ln p). \tag{1.13}$$

Based on the definition of α-entropy, the Renyi dimension is defined as

$$D_\alpha = \lim_{\epsilon\to 0^+} \frac{H_\alpha}{-\ln\epsilon} \tag{1.14}$$

where D_0, D_1 and D_2 are called the capacity, information and correlation dimensions, respectively. In this case, ϵ is a scaling parameter that influences the probabilities p_i.

Chapter 2

Fractal Sets with known Dimension

This chapter summarizes the findings about fractal sets with known dimension and their properties.

2.1 Self Similar Sets

A significant class of fractal set comprises of deterministically self-similar sets. The self similarity guarantees that the part of the set is exactly or approximately the same as the investigated set. Therefore, the set can look the same even when being magnified. To define self similarity, one has to employ the contraction mapping. The mapping $H : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a contraction mapping if for any $\mathcal{F} \subset \mathbb{R}^n$ it holds that

$$\exists q \in (0,1) : \forall \vec{x}, \vec{y} \in \mathcal{F} : \|H(\vec{x}) - H(\vec{y})\| \leqslant q \|\vec{x} - \vec{y}\| . \tag{2.1}$$

The set $\mathcal{F} \subset \mathbb{R}^n$ is self similar [15] iff there is finite number of contraction mappings $\varphi_1, \varphi_2, \ldots, \varphi_k$ such that

$$\mathcal{F} = \bigcup_{i=1}^{k} \varphi_i(\mathcal{F}). \tag{2.2}$$

The Hausdorff dimension of any self-similar set equals to the similarity dimension [15] if the set satisfies the open set condition [7].

A well-known example of self-similar set is Cantor set [16] and n-dimensional Cantor dust [17] with their Hausdorff and Renyi dimension equal to $\log_3 2$ and $n \log_3 2$, respectively. Another famous example of self-similar set is Sierpinski carpet [18] with $D_{\mathrm{H}} = D_\alpha = \log_3 8$.

An example of multifractal, e.g. fractal set with variable Renyi dimension depending on the parameter α is de Wijs' fractal. The graph of De Wijs' fractal [19] with the parameter a is a kind of multifractal that has the Renyi dimension dependent on the dimension parameter α. The corresponding Renyi dimension equals

$$D_\alpha = \frac{1}{1-\alpha} \log_2 \left(a^\alpha + (1-a)^\alpha\right) \tag{2.3}$$

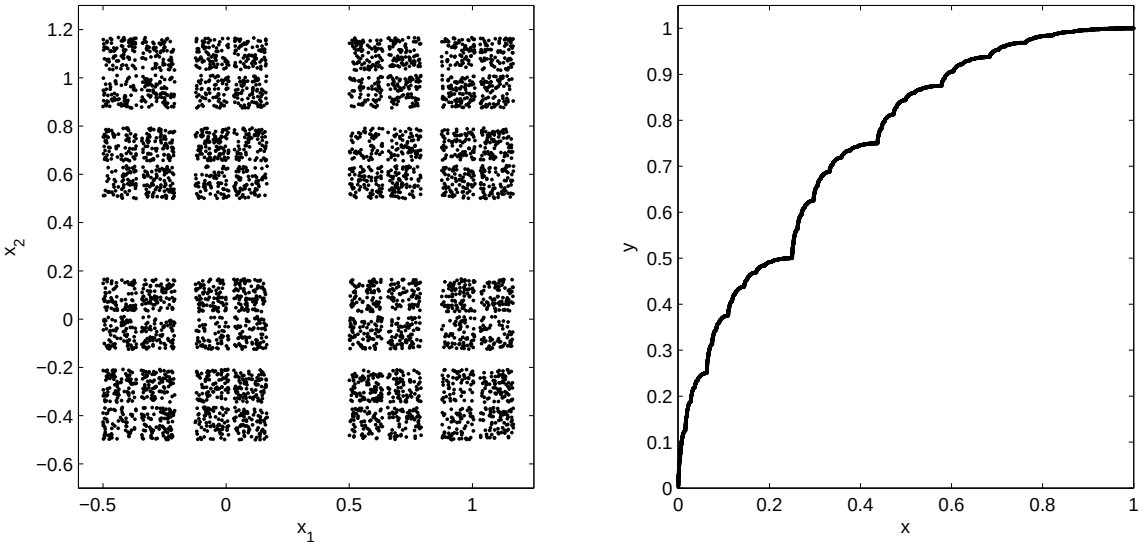


Figure 2.1: Two-dimensional Cantor dust (left) and De Wijs' fractal (right) point sets.

for $0 < a < 1/2$ and $\alpha \in [0; 1) \cup (1, +\infty)$ with the particular case

$$D_1 = \lim_{\alpha \to 1} D_\alpha = -a \log_2 a - (1 - a) \log_2 (1 - a). \tag{2.4}$$

The two dimensional Cantor dust and De Wijs' fractal with $a = 1/4$ are visualised in Fig. 2.1.

2.2 Fractal Functions

There is a class of deterministic functions that have fractal character and noninteger dimension of their graph. The graphs of fractal functions are often self-similar continuous set.

Weierstrass function $W(\vec{x})$ [20] is well-known fractal function defined for any $\vec{x} \in \mathbb{R}^n$ as

$$W(\vec{x}) = \sum_{k=0}^{+\infty} a^k \cos\left(2\pi b^k \left\|\vec{x}\right\|\right). \tag{2.5}$$

where $0 < a < 1$, $b \in \mathbb{N}$ and $ab \geqslant 1$. Weierstrass function is continous in its whole domain, however the derivative doesn't exist anywhere. The Hausdorff dimension of the graph equals [21]

$$D_{\mathrm{H}} = n + 1 + \frac{\log a}{\log b}. \tag{2.6}$$

Takagi function $T(\vec{x})$ [22] has fractal character as well and it is defined as

$$T(\vec{x}) = \sum_{k=0}^{+\infty} a^k \cdot \min_{j=1,..,n} \left\{[2^k x_j], 1 - [2^k x_j]\right\} \tag{2.7}$$

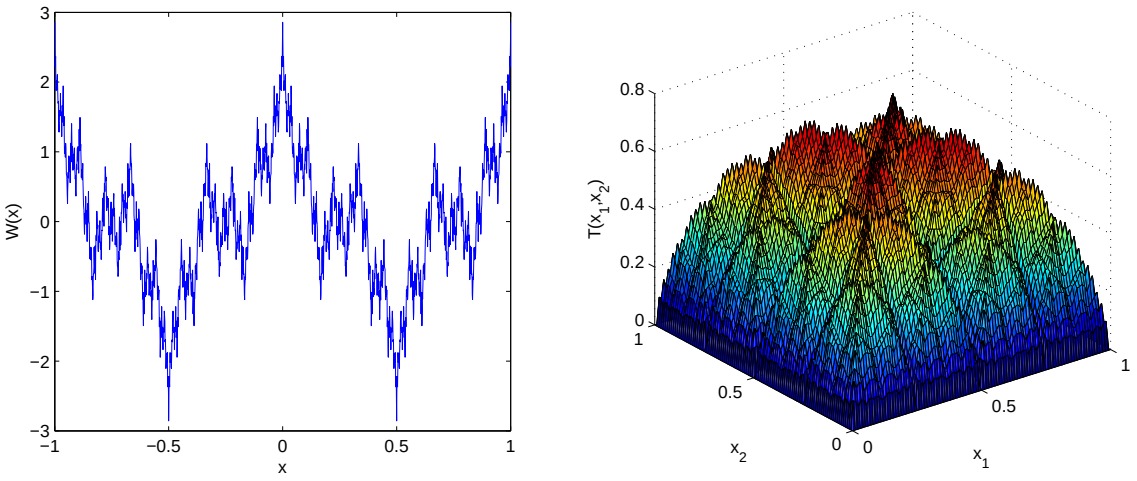


Figure 2.2: One-dimensional Weierstrass (left) and two-dimensional Takagi (right) functions.

for $1/2 < a < 1$ and $[z]$ denotes the fractional part of number $z \in \mathbb{R}$. The Hausdorff dimension of Takagi function equals

$$D_{\mathrm{H}} = n + 1 + \frac{\log a}{\log 2}. \tag{2.8}$$

The one-dimensional Weierstrass function for $a = 0.65$ and $b = 3$ together with Takagi function with parameter $a = 0.55$ are visualized in Fig. 2.2.

2.3 Discrete Dynamical Systems

Discrete dynamic systems with chaotic behaviour generate fractal trajectories and attractors with a nonlinear character. The investigation of this kind of set can be performed in two ways – the first option is to estimate the dimension of the set in the original state space, the second option is to use Whitney's [23, 24] or Takens' [25] theorem and estimate it in a reconstructed space.

Generally, the n-dimensional discrete dynamical process has an internal state $\vec{x}_j \in \mathbb{R}^n$ and output $y_j \in \mathbb{R}$ for $j \in \mathbb{N}_0$. Using reconstruction length $W \in \mathbb{N}$, we define a sliding sample $\vec{\xi}_j = (y_j, \ldots, y_{j+W-1}) \in \mathbb{R}^W$ for $j \in \mathbb{N}_0$, first. Whitney,s embedding theorem can be rewritten from continuous to discrete time as follows: When $W \geqslant 2n + 1$, then the reconstructed series $\{\vec{\xi}_j\}_{j=0}^{\infty}$ has the same properties as $\{\vec{x}_j\}_{j=0}^{\infty}$. Therefore, any Renyi dimension D_α of the reconstructed attractor is the same as in the case of the state space.

The Henon map is two-dimensional discrete dynamical process defined by the following set of equations

$$x_{n+1} = 1 - ax_n^2 + y_n, \tag{2.9}$$
$$y_{n+1} = bx_n, \tag{2.10}$$

Table 2.1: Attractor dimensions for discrete dynamical systems.

System	D_0	D_1	D_2
Henon map [26, 27]	1.261	1.2583	1.2201
Lozi map [28, 29]	unknown	1.4042	1.3845

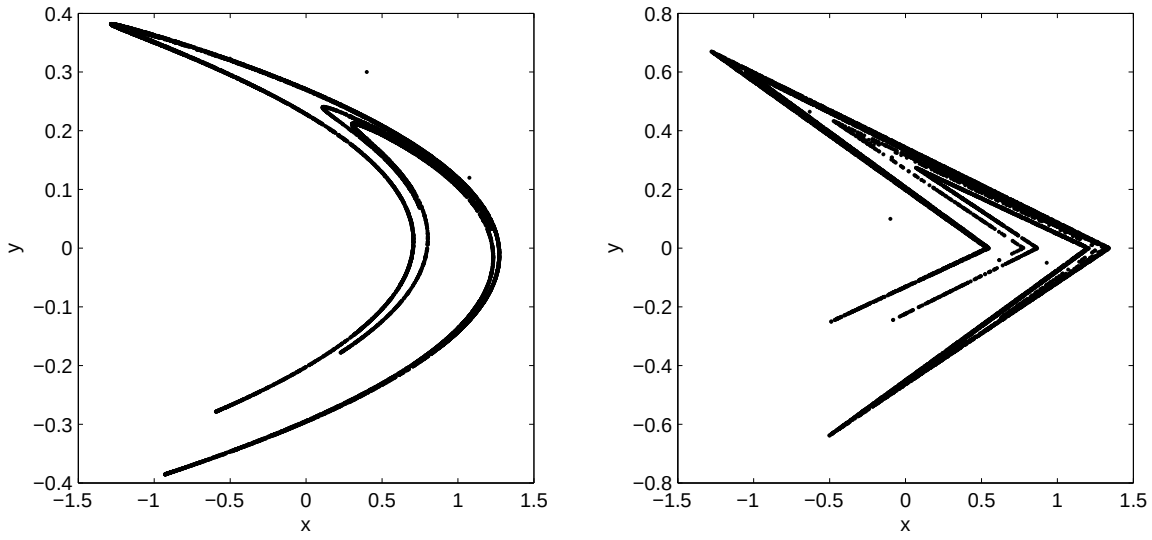


Figure 2.3: Trajectories of Henon (left) and Lozi (right) maps.

while Lozi map uses the following definition

$$x_{n+1} = 1 - a|x_n| + y_n, \tag{2.11}$$
$$y_{n+1} = bx_n. \tag{2.12}$$

The information and correlation dimension is estimated with high precision for the typical representatives of discrete dynamic systems. Tab. 2.1 shows the Renyi dimension for Henon map with the parameters $a = 0.4, b = 0.3, x_0 = 0, y_0 = 0.9$ and Lozi map with the parameters $a = 1.7, b = 0.5, x_0 = -0.1, y_0 = 0.1$. The systems for the given parameters are visualized in Fig. 2.3.

2.4 Fractional Processes

Fractional Brownian motion (fBm) [2] is a continuous Gaussian process $B_H(t)$ defined for $t \in [0; +\infty)$, $H \in (0; 1)$ and $\sigma > 0$. The process starts at zero and has zero expected value for all positive times t. The autocovariance structure of fBm obeys for all $t, s > 0$

$$\mathrm{E}(B_H(t)B_H(s)) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right), \tag{2.13}$$

where E is mean value operator. Parameter H is called Hurst exponent and it is related to the Hausdorff dimension of fBm trajectory that equals $D_{\mathrm{H}} = 2 - H$. Moreover, any Rényi dimension D_α [30] has the same value as D_{H} for $\alpha \geqslant 0$. The parameter σ is often normalized to be unit and we will denote this special case as standardized fractional Brownian motion. For $H = 1/2$, the standard fBm becomes Wiener process $W(t)$, which is standard Brownian motion.

Fractional Gaussian noise (fGn) is the process $G_H(t)$ defined for all $t > 0$ as

$$G_H(t) = B_H(t+1) - B_H(t). \tag{2.14}$$

However, the Hausdorff and Renyi dimension of fGn is independent of H and equals always $D_{\mathrm{H}} = 2$. The process is still Gaussian, zero mean and if it is constructed on the basis of standardized fBm, it has also unit variance and the autocorrelation function can be expressed as a function of $k \in \mathbb{Z}$ and $H \in (0;1)$ in the form

$$\eta(k,H) = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right). \tag{2.15}$$

Another derived process from the fBm is called Brownian bridge. The Brownian bridge (BB) $M(t)$ is defined using standard Wiener process $W(t)$ on the unit interval as

$$M(t) = W(t) - tW(1). \tag{2.16}$$

BB is again Gaussian process with stationary, but not independent increments. From the definition, the value in both endpoints equals zero i.e $M(0) = M(1) = 0$, and the autocovariance structure is given by

$$\mathrm{E}(M(t)M(s)) = \min(t,s) - ts. \tag{2.17}$$

for $t, s \in (0,1)$.

2.5 Multivariate Random Processes

Levy flight [31] $L_\alpha(t)$ is a special class of multidimensional random processes. It is a group of Markov processes on the interval $T = [0;\infty)$ with increments from distribution P, e.g.

$$\|L_\alpha(t+1) - L_\alpha(t)\| \sim \mathrm{P} \tag{2.18}$$

satisfying

$$1 - \mathrm{P}(u) = \mathcal{O}(u^{-\alpha}) \tag{2.19}$$

for any $0 < \alpha < 2$ and $u \to \infty$. When the process is realized in n-dimensional space and when the point set $\mathcal{F} \subset \mathbb{R}^n$ consists only of its dependent values

$$\mathcal{F} = \{L_\alpha(t) \in \mathbb{R}^n : t \in [0;\infty)\}, \tag{2.20}$$

the Hausdorff dimension equals [32] $D_{\mathrm{H}} = \min\{\alpha, n\}$.

Fractional Brownian trace $T_H(t)$ is n-dimensional process defined for any $t \in \mathbb{R}$ as

$$T_H(t) = \left(B_H^1(t), B_H^2(t), \ldots, B_H^n(t)\right), \tag{2.21}$$

where $B_H^i(t)$ are one-dimensional independent fBms for every $i = 1, \ldots, n$. The Hausdorff dimension of $\mathcal{F} \subset \mathbb{R}^n$ consisting only of its dependent values as

$$\mathcal{F} = \{T_H(t) \in \mathbb{R}^n : t \in [0; \infty)\}, \tag{2.22}$$

equals

$$D_{\mathrm{H}} = \min\left\{\frac{1}{H}, n\right\}. \tag{2.23}$$

The one-dimensional fBm with Hurst exponent $H \in (0, 1)$ and variance $\sigma^2 > 0$ can be generalized also in another way [33, 34]. Let $\vec{x} \in \mathbb{R}^n$ be independent variable of dimension $n \in \mathbb{N}$. The n-dimensional fBm [35] denoted as $B_H(\vec{x})$ satisfies

$$\mathrm{E}\left(B_H(\vec{x})B_H(\vec{y})\right) = \frac{\sigma^2}{2} \cdot \left(\|\vec{x}\|^{2H} + \|\vec{y}\|^{2H} - \|\vec{x} - \vec{y}\|^{2H}\right) \tag{2.24}$$

and $\mathrm{E}(B_H(\vec{x})) = 0$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $\sigma > 0$,where $\|..\|$ denotes Euclidean distance. The Hausdorff and Renyi dimension of graph of n-dimensional fBm equals [36] $D_{\mathrm{H}} = D_\alpha = n + 1 - H$.

Chapter 3

Fractal Dimension Estimation

3.1 Grid Technique

Grid technique is a method of data gathering that allows to place any bounded finite-dimensional point sets in $\mathbb{R}^n$ into a predefined grid structure. The method involves general and systematic point placement from the investigated set into a regular structure that is formed by n-dimensional hypercubes with side size $\epsilon > 0$.

For a bounded set $\mathcal{F} \subset \mathbb{R}^n$ and for every $i \in \{1, \ldots, n\}$, we define the lower and upper bound in each dimension as

$$u_i = \max\{x_i : \vec{x} \in \mathcal{F}\} \tag{3.1}$$

$$l_i = \min\{x_i : \vec{x} \in \mathcal{F}\}. \tag{3.2}$$

The set of possible grid indexes is defined as

$$\mathcal{S}_\epsilon = \bigtimes_{1 \leqslant i \leqslant n} \left\{ 0, 1, \ldots, \left\lceil \frac{u_i - l_i}{\epsilon} \right\rceil \right\}, \tag{3.3}$$

therefore we can define a hypercube in the grid for any $\vec{k} \in \mathcal{S}_\epsilon$ as

$$\mathcal{A}(\vec{k}) = \bigtimes_{1 \leqslant i \leqslant n} \left[l_i + k_i\epsilon, l_i + (k_i + 1)\epsilon \right). \tag{3.4}$$

We define partial ordering on the set $\mathcal{S}_\epsilon$ as follows. Let $\vec{x}, \vec{y} \in \mathcal{S}_\epsilon$, then

$$\vec{x} \preccurlyeq \vec{y} \qquad \text{iff} \qquad \sum_{i=1}^{n} 10^p x_{n-i+1} \leqslant \sum_{i=1}^{n} 10^i y_{n-i+1}. \tag{3.5}$$

Denote the total number of hypercubes as

$$m = \prod_{i=1}^{n} \left\lceil \frac{u_i - l_i}{\epsilon} \right\rceil + 1. \tag{3.6}$$

Using the ordering (3.5) we can define the grid $\mathcal{G}_\epsilon$ as a sequence of all hypercubes that are defined by equation (3.4), i.e.

$$\mathcal{G}_\epsilon = (\mathcal{A}(\vec{k}_1), \mathcal{A}(\vec{k}_2), \dots \mathcal{A}(\vec{k}_m)) \tag{3.7}$$

where $\vec{k}_i \preccurlyeq \vec{k}_{i+1}$ for $i = 1, \dots m-1$ and $k_i \in \mathcal{S}_\epsilon$ for $i = 1, \dots m$. For simplicity, we will later denote the grid as

$$\mathcal{G}_\epsilon = (\mathcal{A}_i)_{i=1}^m. \tag{3.8}$$

Analysing data via grid technique has several advantages. Most of the time we will be interested in the number of hypercubes containing at least one point of the investigated set $\mathcal{F} \subset \mathbb{R}^n$. The number of hypercubes containing at least one point from $\mathcal{F}$ is denoted as N_ϵ is defined as

$$N_\epsilon = \operatorname{card}\left\{\mathcal{A}(\vec{k}) : \vec{k} \in \mathcal{S}_\epsilon \text{ and } \mathcal{F} \cap \mathcal{A}(\vec{k}) \neq \varnothing\right\}. \tag{3.9}$$

3.2 Random Sampling

Given measurable set $\mathcal{F} \subset \mathbb{R}^n$ a random sample of length $N \in \mathbb{N}$ is a set of points

$$\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N \sim \mathrm{U}(\mathcal{F}). \tag{3.10}$$

Most well-known fractal sets can be

- a self-similar set defined using translation rules with infinite steps (e.g. Cantor set),
- a fractional function defined on its domain using a prescribed formula,
- a fractional process defined using its autocorrelation function.

Most self-similar sets can be defined using infinite subdivision rule. In each iteration, the structure replicates itself, diminishes in size and moves towards a particular direction. Knowing all possible directions, their probability and scale change, one can start generating the points from such set by starting at predefined point in $\mathbb{R}^n$. At every iteration, shifting this point in one of the directions and scaling the direction vector, results in a sample generated in certain depth. The Fig. 3.1 shows the sampling of the Sierpinski carpet for 1000 and 10 000 points, respectively.

Taking fractional functions into account, we introduce a methodology for generating samples from graph of continuous function with a bounded domain. An initial idea of generating points uniformly from the domain and completing their functional values will not result in uniform sampling from its graph. In one dimension, the procedure can be very straightforward – we first generate uniformly the points from the function's domain and sort them in ascending order. Connecting their functional values using abscissas in the respective order will result in a sample trajectory in

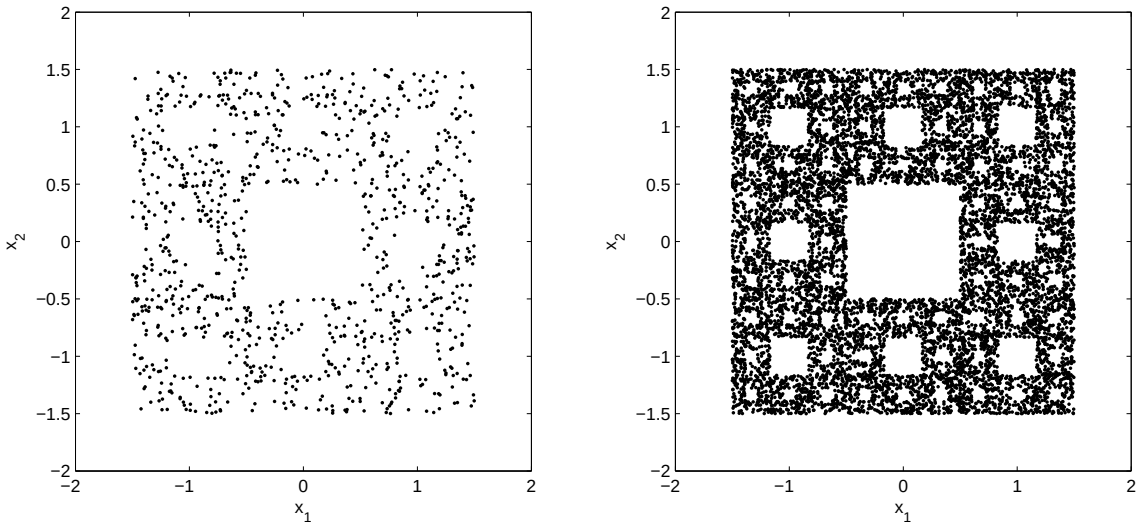


Figure 3.1: Sierpinski carpet sample of 1000 points (left) and 10 000 points (right).

$\mathbb{R}^2$. Generating uniformly points from the trajectory will assure a uniform sample from the graph. An example of Takagi function is in Fig. 3.2.

The procedure for generation fractional processes with given autocorrelation function is thoroughly described in section 3.7.

3.3 Methodological Aspects of Dimension Estimation

When estimating the fractal dimension or Hurst exponent of fractal sets, the bias is often present in the estimation. No matter how detailed structure we have, the point estimate can be biased, since the conclusion comes from finite set of points, whereas chaotic systems always consist of uncountable number of points. Therefore, even the sampling is random, the estimation error is always present. Usually, the bigger the sample, the more details of fractal patterns in the structure can be found and the lower the bias, however it is never possible to fully get rid of it.

To anticipate against huge bias, the methods are always assuming one or more theoretical properties to hold in order to present more accurate estimation. Thorough this chapter, we introduce box-counting method for D_0 estimation, correlation sum for D_2 estimation and general Renyi method for D_α estimation.

3.4 Box-Counting Method

Box-counting is a grid method that supposes the placement of a sample of a fractal set $\mathcal{F} \subset \mathbb{R}^n$ into a grid that is formed by n-dimensional hypercubes with side size

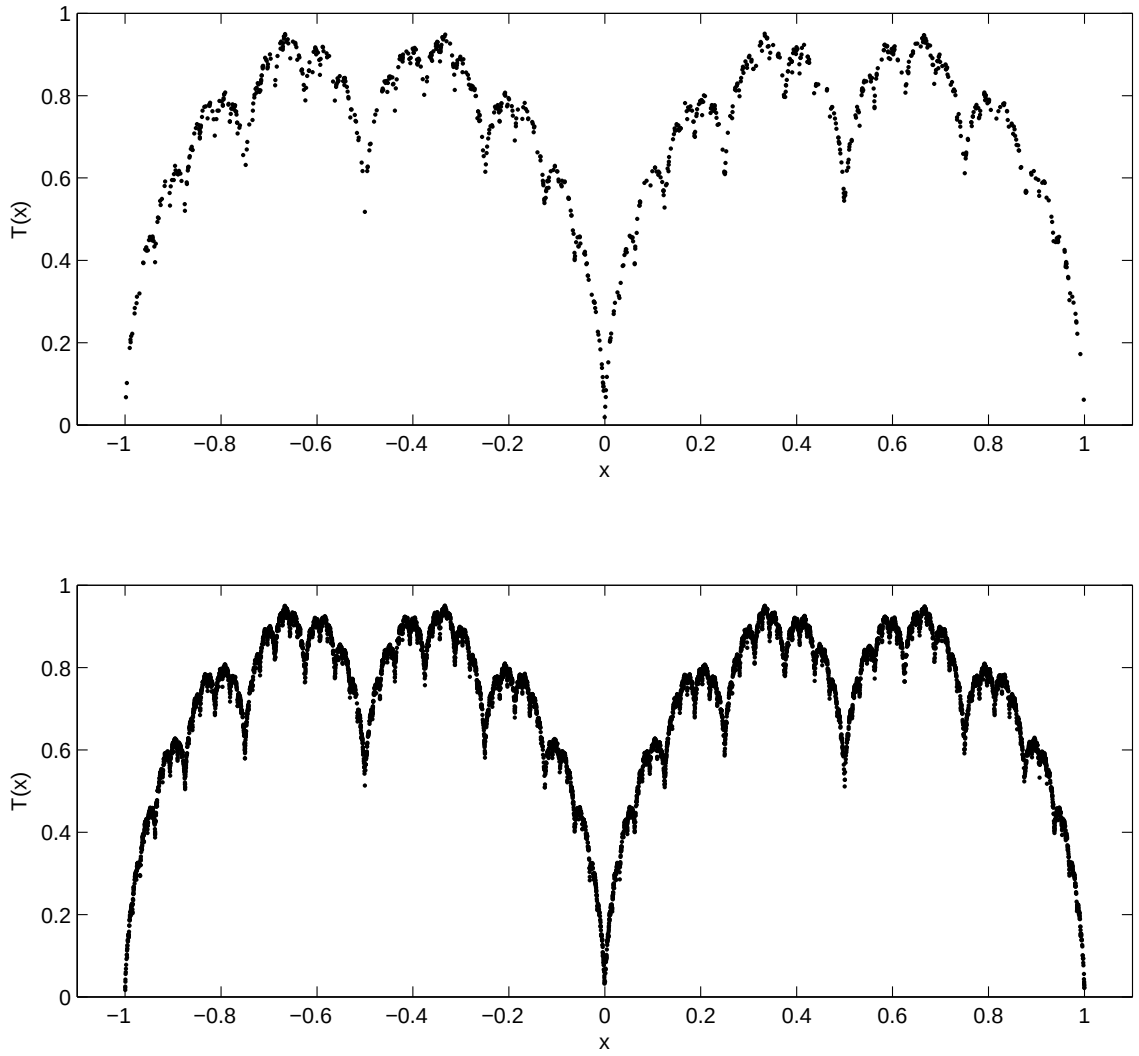


Figure 3.2: Takagi function graph sample of 1000 points (top) and 10 000 points (bottom).

$\epsilon > 0$. The method works under the assumption that the number of hypercubes $N_\epsilon(\mathcal{F})$ containing at least one point of the sample follows the power law [3]

$$N_\epsilon(\mathcal{F}) \propto \epsilon^{-D_0}, \tag{3.11}$$

where D_0 is the capacity dimension of $\mathcal{F}$. For any non-empty $\mathcal{F} \subset \mathbb{R}^n$ it holds [3] that

$$D_0 = \lim_{\epsilon \to 0^+} \frac{\ln N_\epsilon(\mathcal{F})}{-\ln \epsilon}, \tag{3.12}$$

if the limit on the right side of the equation exists and $N_\epsilon(\mathcal{F})$ denotes the number

of hypercubes from the grid that have non-empty intersection with the fractal set. The capacity dimension is estimated using linear regression model

$$\ln N_\epsilon(\mathcal{F}) = A - D_0 \cdot \ln \epsilon \tag{3.13}$$

for small values of parameter ϵ.

3.5 Correlation Sum Evaluation

Recall that the correlation dimension is based on the calculation of Euclidean distance between points in set $\mathcal{F} \subset \mathbb{R}^n$ and the correlation sum $C(r)$ is defined for any $r > 0$ as

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I\ (\|\vec{x_i} - \vec{x_j}\| \leqslant r) \tag{3.14}$$

for $\vec{x_i}, \vec{x_j} \in \mathcal{F}$ a indicator function I.

Recall the definition of correlation dimension as

$$D_2 = \lim_{r\to 0^+} \frac{\ln\ C(r)}{\ln\ r}, \tag{3.15}$$

if the limit on the right side of the equation exists. If the investigated set $\mathcal{F} \subset \mathbb{R}^n$ has finite number of points, the correlation dimension always equals zero. The estimation of correlation dimension from a finite sample is again performed using linear regression model

$$\ln\ C(r) = A + D_2 \cdot \ln\ r \tag{3.16}$$

for small values of r.

3.6 Renyi Entropy and Dimension Estimation

Recall that the Renyi entropy is defined for $\alpha \geqslant 0$ as

$$H_\alpha(\vec{p}) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{k} p_i^\alpha \tag{3.17}$$

for $\alpha \neq 1$ and

$$H_1 = \lim_{\alpha\to 1} H_\alpha = -\sum_{i=1}^{k} p_i \ln p_i \tag{3.18}$$

where k is the number of events and p_i are their probabilities satisfying $\sum p_i = 1$.

Let $\mathcal{A}_i \subset \mathbb{R}^n$ be grid element as hypercube of size ϵ for $i = 1, ..., k$ satisfying $\mathcal{F} \subset \bigcup_{i=1}^{k} \mathcal{A}_i$. Recall the system of non-overapping elements is denoted as a grid

$\mathcal{G}_\epsilon = (\mathcal{A}_i)_{i=1}^k$. The probabilities p_i are calculated [37, 38] as

$$p_i = \frac{\mu(\mathcal{A}_i \cap \mathcal{F})}{\mu(\mathcal{A}_i)}, \tag{3.19}$$

for $i = 1, \ldots, k$, where $\mu(...)$ denotes the volume of the set. The variable p_i expresses the probability that a randomly taken point from set $\mathcal{F}$ will be contained in the grid box $\mathcal{A}_i$. Having a finite sample, the probabilities are approximated as the number of points that are contained in a particular box divided by the total number of points in the sample.

Recall the definition of the Renyi dimension as

$$D_\alpha = \lim_{\epsilon \to 0^+} \frac{H_\alpha}{-\ln \epsilon}, \tag{3.20}$$

the estimation of the Renyi dimension from finite sample can be calculated via linear regression model

$$H_\alpha = A - D_\alpha \cdot \ln\, \epsilon \tag{3.21}$$

for small values ϵ.

3.7 Simulation and Parameter Estimation of fBm

There is a variety of methods for fBm sample generation with predefined Hurst exponent H. The most accurate ones focus on drawing the resulting time-series according to it's autocovariance function. Lowen method [39] relies on the spectral power in the frequency domain. At first, the function $R_x(l)$ is defined recursively with period $2N$ as

$$R_x(l) = \begin{cases} \frac{1}{2}(1 - \frac{l}{N})^{2H} & : 0 \leqslant l \leqslant N \\ R_x(2N - l) & : N < l \leqslant 2N \end{cases}.$$

Based on it's spectral power density S_x, we define

$$X(f) = \begin{cases} 0 & f = 0 \\ \exp(i\theta_k) g_k \sqrt{S_x(f)} & 0 < f < N \\ g_k \sqrt{S_x(f)} & f = N \\ X^*(2N - f) & N < f < 2N \end{cases},$$

where * denotes the complex conjugate, $\theta_k \sim \langle 0; 2\pi)$ and $g_k \sim \mathrm{N}(0, 1)$. The resulting sample is obtained as inverse Fourier transformation [40] of $X(f)$. The only limitation of Lowen method is that it only works for $0 < H < 0.5$ [39]. Another technique of fBm generation is Abry–Sellan method [41]. This method is based on integral representation of fBm and can be expressed as a convergent sum [42] using wavelets as

$$B_H(t) = \sum_{k=-\infty}^{+\infty} \Phi_H(t - k) S_k^{(H)} + \sum_{j=0}^{+\infty} \sum_{k=-\infty}^{+\infty} 2^{-jH} \Psi_H(2^j t - k) g_{j,k}, \tag{3.22}$$

where Φ_H is bi-orthonormal scaling function [43], Ψ_H is wavelet, $S_k^{(H)}$ is fractional ARFIMA $(0, H-1/2, 0)$ process [44] and $g_{j,k}$ are sampled iid from normal distribution. The method has limited range and reliably works for $H > 1/2$ [45]. A general approach for generating any (non)stationary Gaussian process with specified autocorrelation structure is represented by Davies–Harte method [46]. For given autocovariance coefficients $c_0, c_1, \ldots, c_n$ and process of length $n+1$ we define

$$g_k = \sum_{j=0}^{n-1} c_j \exp\{2\pi \mathrm{i} kj/(2n)\} + \sum_{j=n}^{2n-1} c_{2n-j} \exp\{2\pi \mathrm{i} kj/(2n)\} \tag{3.23}$$

When Z_k is a sequence of independent complex normal random variables [46] that have independent real and imaginary parts for $0 \leqslant k \leqslant n$, $Z_k = \overline{Z}_{2n-k}$ for $n < k < 2$. Coefficients Z_0 and Z_n have variance two, while Z_k for $0 < k < n$ has unit variance for both real and imaginary part. By defining

$$x_j = \frac{1}{2} n^{-\frac{1}{2}} \sum_{k=0}^{2n-1} Z_k g_k^{1/2} \exp\{2\pi \mathrm{i} jk/(2n)\} \tag{3.24}$$

for $0 \leqslant j \leqslant k$, the time series $\{x_j\}$ has the required distribution [47]. Simple, but powerful method for generating fBm sequences is represented by circulant embedding method [48]. This approach is the generalization of traditional Cholesky algorithm and allows to generate longer sequences with lower computational memory and time complexity [49].

As the generation methods of fBm trajectories are known, we can introduce a few traditional methods that serve for Hurst Exponent estimation.

The R/S method, a classical technique of Hurst exponent [1] estimation, repeatedly separates the original sampling of a signal into disjoint segments and calculates mean, standard deviation and range for each of them. Considering the division into $r \in \mathbb{N}$ segments with $m \in \mathbb{N}$ elements, the R/S statistic is given by

$$(R/S)_m = \frac{1}{r} \sum_{j=1}^{r} \frac{R_j}{S_j}, \tag{3.25}$$

where S_j and R_j indicates standard deviation and range at interval j, respectively. For different interval lengths m, the R/S statistics $(R/S)_m$ is subsequently plot against m in log-log plot to obtain Hurst parameter using the power law

$$\mathrm{E}\left[(R/S)_m\right] \propto m^H. \tag{3.26}$$

Zero-crossing method, independently investigated by Azais [50], Feuerverger [51] and Couerjolly [48] provides point estimate of Hurst parameter H based on the number of zero-crossing of fGn sample. Couerjolly suggests to calculate the relative number of zero crossing S_N with respect to the sample length N, which converges almost surely to function $\theta(H)/\pi$, where $\theta(H)$ is defined as

$$\theta(H) = \begin{cases} \operatorname{atan}\left(\frac{\sqrt{1-r(H)^2}}{r(H)}\right) & \text{for } r(H) > 0 \\ \frac{\pi}{2} + \operatorname{atan}\left(\frac{-r(H)}{\sqrt{1-r(H)^2}}\right) & \text{otherwise} \end{cases}$$

and $r(H) = 2^{2H-1}-1$. For finite N large enough, the Hurst parameter approximately equals

$$H \approx \frac{1}{2}\left(1 + \log_2(1 + \epsilon|\cos(\pi \cdot S_N)|)\right), \tag{3.27}$$

where $\epsilon = \operatorname{sgn}\left(H - \frac{1}{2}\right)$ is assumed to be known.

Whittle estimator [52] is a method for estimation of fractional parameter d of ARFIMA(p, d, q) process. The covariance function of fGn with $H = d + \frac{1}{2}$ obeys the same power law as ARFIMA(0,d,0), however their spectral densities differ. The aim is to estimate the parameter d_{opt} by solving the minimization problem

$$d_{\text{opt}} \in \underset{d \in \left(-\frac{1}{2};\frac{1}{2}\right)}{\operatorname{argmin}} Q(d) \quad \text{for} \quad Q(d) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda, d)} \mathrm{d}\lambda, \tag{3.28}$$

where $I(\lambda)$ is the experimental spectral density of investigated sample and $f(\lambda, d)$ is theoretical density of ARFIMA(0,d,0) in the form

$$f(\lambda, d) = \frac{1}{2\pi}\left(2\sin\frac{\lambda}{2}\right)^{-2d}. \tag{3.29}$$

Istas and Lang [53] designed a method for Holder parameter estimation of random process. Since fBm is Holder continuous with parameter $H - \epsilon$ for every $\epsilon > 0$, the upper estimate of it can lead to a good estimation of Hurst exponent H. For a sample of random process $X(i), i = 0, ..., n$ and arbitrary vector $\vec{a} \in \mathbb{R}^{p+1}$ satisfying the condition, that sum of its elements equals zero, they defined quadratic variation with sampling step Δ

$$U(\vec{a}, n, \Delta) = \frac{1}{n}\sum_{j=0}^{n-p} (\Delta_{\vec{a}} X_j)^2 \quad \text{where} \quad \Delta_{\vec{a}} X_j = \sum_{i=0}^{p} a_i X((i + j) \cdot \Delta). \tag{3.30}$$

Istas and Lang proved, that the quadratic variation $U(\vec{a}, n, \Delta)$ converges to variance $\sigma^2_{\vec{a},\Delta}$ of finite difference $\Delta_{\vec{a}} X_j$ as n tends to infinity. The variance can be expressed by following formula

$$\sigma^2_{\vec{a},\Delta} \approx -C \cdot (-1)^D \sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l |k - l|^{2H} \cdot \Delta^{2H}, \tag{3.31}$$

for constants $C \in \mathbb{R}$, $D \in \mathbb{N}$ and Hurst exponent H. It is possible to estimate the parameter H by utilizing different choices of $\vec{a}$ for nonlinear regression model (3.31).

Chapter 4

Thesis Aims

This thesis is extending the known theory of fractal sets and dimension estimation in the area of point sets and fractional processes parameter estimation. There are the aims that were the subject of this thesis.

Develop new method for correlation dimension estimation. Currently the commonly used methodology is the technique of correlation sum, that is sufficient for many real world applications, but it comes with a significant bias. Developing a new method that could provide estimation with the same precision/variance and remove the bias would be beneficial for theoretical field as well as practical applications.

Develop new method for Renyi dimension estimation. Currently, there are methods for estimating capacity, information and correlation dimension, however there is lack of general methods that could estimate any Renyi dimension D_α even for non-integer parameters α. This would allow to analyze variety of fractal structures including multifractals.

Develop new method for Hurst exponent estimation. Nowadays, the fractal investigation of signal often relies just on R/S method or spectral approach that don't reliably work for extreme values of Hurst exponent. The aim is to propose approaches for Hurst exponent estimation to correctly analyze short time series as well as multidimensional fractional surfaces and masses. These approaches should always aim on diminishing the variance and reducing bias.

Apply the methods to artificial sets with known dimension. The proposed novel methods shall be applied to sets and signal samples with known dimension that can serve as benchmark of the estimation quality and unbiasedness. There are variety of artificial sets that can serve well for extensive testing.

Apply the methods to biomedical and econometric signals. Once the methods are validated on sets with known dimension, the techniques can be advantageously used for analysis of biomedical and econometric signals. There exists evidence that biomedical data such as EEG or mammography exhibit chaotic behaviour, thus represents a suitable example to model on using the developed methods.

Chapter 5

Spectral Approach to D_2 Estimation

This chapter introduces a new method for estimation of D_2 (correlation dimension). We introduce a new characteristics that has smoother development than traditional correlation sum and prove its properties to verify it is suitable for correlation dimension estimation. This method is a spectral method, e.g. based on a power spectrum of a set, which has for most self similar sets not smooth dependence on frequency.

The correlation dimension of fractal set can be investigated using the power spectrum of the set. The Fourier transform of an n-dimensional set $\mathcal{F} \subset \mathbb{R}^n$ is defined using the operator of the expected value [54] as

$$F(\vec{\omega}) = \underset{\vec{x}\sim \mathrm{U}(\mathcal{F})}{\mathrm{E}} \exp(-\mathrm{i}\vec{\omega} \cdot \vec{x}) \tag{5.1}$$

for angular frequency $\vec{\omega} \in \mathbb{R}^n$ and for $\vec{x}$ uniformly distributed on $\mathcal{F}$. The traditional continuous power spectrum of set $\mathcal{F} \subset \mathbb{R}^n$ is defined as

$$P(\vec{\omega}) = \underset{\vec{x}\sim \mathrm{U}(\mathcal{F})}{\mathrm{E}} \underset{\vec{y}\sim \mathrm{U}(\mathcal{F})}{\mathrm{E}} \exp(-\mathrm{i}\vec{\omega} \cdot \vec{x}) \exp(\mathrm{i}\vec{\omega} \cdot \vec{y}) = \underset{\vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})}{\mathrm{E}} \exp(-\mathrm{i}\vec{\omega}(\vec{x} - \vec{y})). \tag{5.2}$$

The power spectrum does not have suitable properties for statistical analysis. Most often, we suppose the exponential dependency of power spectrum $P(\vec{\omega})$ on the correlation dimension D_2 as

$$P(\vec{\omega}) \propto \omega^{-D_2}, \tag{5.3}$$

which can lead to inaccurate results due to the spectrum instability for the majority of self-similar sets. The power spectra of Cantor set is visualized in Fig. 5.1.

5.1 Physical Motivation of Spectral Method

The aim of this approach is to smoothen the power spectra by terms of rotating it around the origin. The goal of the novel method is to obtain a one-dimensional function as a derivative of the power spectrum, which is useful in fractal analysis. The procedure was inspired by Debye [55] and by his X-ray diffraction method, which is often referred to as the Debye-Scherrer method. We denote $\mathrm{SO}(n)$ as the

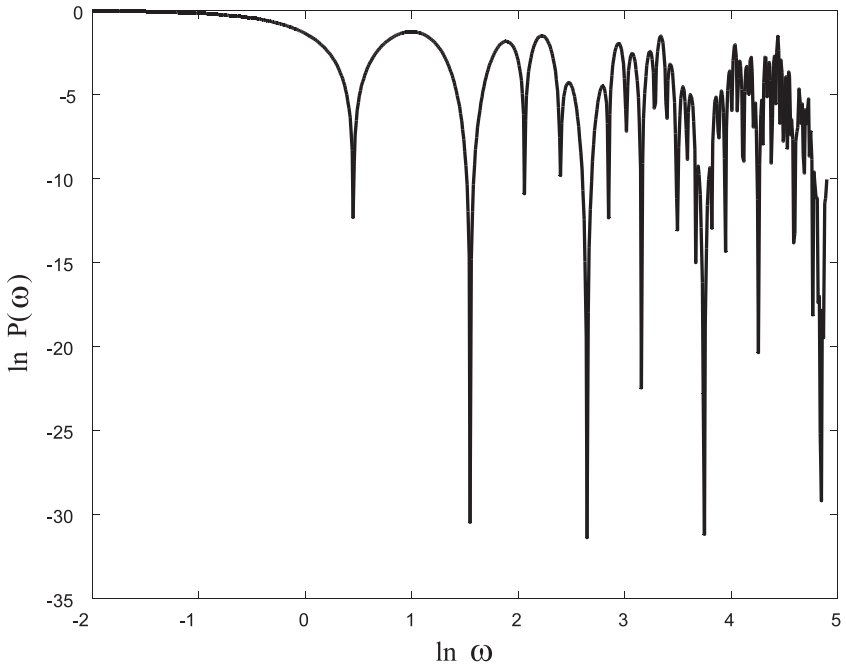


Figure 5.1: Power spectrum of one-dimensional Cantor set.

group of all rotations in $\mathbb{R}^n$ around the origin. Because any rotation $R \in \mathrm{SO}(n)$ is a linear transform, the following equation holds

$$R(\vec{x}) - R(\vec{y}) = R(\vec{x} - \vec{y}) = \|\vec{x} - \vec{y}\| \cdot \vec{\xi}, \tag{5.4}$$

where $\vec{\xi}$ is a direction vector satisfying $\left\|\vec{\xi}\right\| = 1$ and $\vec{\xi} \in \mathcal{S}_{n-1}$ for an n-dimensional sphere $\mathcal{S}_{n-1} = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\| = 1\}$. Using the factorisation of angular frequency $\vec{\omega} = \Omega \cdot \vec{\psi}$ for $\Omega \in \mathbb{R}_0^+$ and normalisation vector $\vec{\psi} \in \mathcal{S}_{n-1}$, we can define rotational spectrum as

$$S(\Omega) = \mathop{\mathrm{E}}_{R \in \mathrm{SO}(n)} \mathop{\mathrm{E}}_{\vec{\psi} \in \mathcal{S}_{n-1}} \mathop{\mathrm{E}}_{\vec{x},\vec{y} \sim U(\mathcal{F})} \exp(-\mathrm{i}\Omega \vec{\psi} R(\vec{x} - \vec{y})), \tag{5.5}$$

which can be expressed explicitly in the following theorem.

Theorem 1. *Rotational spectrum can be expressed as*

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y} \sim U(\mathcal{F})} H_n(\Omega \|\vec{x} - \vec{y}\|), \tag{5.6}$$

where

$$H_n(q) = \frac{2^{\frac{n-2}{2}} \cdot \Gamma\left(\frac{n}{2}\right)}{q^{\frac{n-2}{2}}} J_{\frac{n-2}{2}}(q). \tag{5.7}$$

Proof. Because every rotation is a linear transform, we can rewrite the rotational spectrum as

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y} \sim U(\mathcal{F})} \mathop{\mathrm{E}}_{\vec{\psi},\vec{\xi} \in \mathcal{S}_{n-1}} \exp(-\mathrm{i}\Omega \|\vec{x} - \vec{y}\| \, \vec{\psi} \cdot \vec{\xi}). \tag{5.8}$$

The angle ν between vectors $\vec{\psi}$ and $\vec{\xi}$ satisfies $\cos\nu = \vec{\psi}\cdot\vec{\xi}$. Without the loss of generality, we can set $\vec{\xi} = (1, 0, 0, \ldots, 0)$ and rewrite the rotational spectrum as

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y}\in\mathcal{F}} \quad H_n\left(\Omega\|\vec{x}-\vec{y}\|\right), \tag{5.9}$$

where the function $H_n : \mathbb{R} \mapsto \mathbb{C}$ is defined as

$$H_n(q) = \mathop{\mathrm{E}}_{\substack{\vec{\psi}\in\mathcal{S}_{n-1}\\ \psi_1=\cos\nu}} \quad \exp(-\mathrm{i}q\cos\nu). \tag{5.10}$$

For $n = 1$, we obtain a degenerated rotation together with $\nu \in \{0, \pi\}$; therefore, the kernel function H_1 equals

$$H_1(q) = \frac{\exp(-\mathrm{i}q) + \exp(\mathrm{i}q)}{2} = \cos q. \tag{5.11}$$

In case $n \geqslant 2$, we can express the kernel function using an integral formula:

$$H_n(q) = \frac{I_1(q)}{I_2(q)} = \frac{\int_0^\pi \exp(-\mathrm{i}q\cos\nu)\sin^{n-2}\nu\,\mathrm{d}\nu}{\int_0^\pi \sin^{n-2}\nu\,\mathrm{d}\nu}. \tag{5.12}$$

The Poisson integral [56] formula for the Bessel function [57] $J_p(q)$ of the first kind in the form

$$J_p(q) = \frac{\left(\frac{q}{2}\right)^p}{\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}\int_0^\pi \exp(-\mathrm{i}q\cos\nu)\sin^{2p}\nu\,\mathrm{d}\nu \tag{5.13}$$

allows the integral in the nominator to be rewritten as

$$I_1(q) = \frac{J_p(q)\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\left(\frac{q}{2}\right)^p}, \tag{5.14}$$

whereas the integral in the denominator is a limit case of the Poisson formula

$$I_2(q) = \lim_{q\to 0} \quad \frac{J_p(q)\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\left(\frac{q}{2}\right)^p} = \frac{\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\Gamma\left(p+1\right)}. \tag{5.15}$$

For $p = \frac{n-2}{2}$, we obtain the final form of the kernel function expressed by the Bessel function $J_p(q)$ as

$$H_n(q) = \frac{2^{\frac{n-2}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}{q^{\frac{n-2}{2}}}J_{\frac{n-2}{2}}(q). \tag{5.16}$$

Applying $H_n(q)$ for $n = 1$, we obtain $H_1(q) = \cos q$ as a particular case, which extends the range of formula (5.16) to $n \in \mathbb{N}$ and completes the proof. □

The rotation can be performed in any space that has dimension n not smaller than the dimension m of the original space of $\mathcal{F}$. When the dimension of the rotation is greater than m, any vector $\vec{x} \in \mathcal{F}$ is completed with the zeros for the remaining $n-m$ coordinates having a sufficient length. The most valuable result can be obtained in the case of rotation in an infinite-dimensional space.

Theorem 2. *The scaled limit case of the kernel function H_n is the Gaussian function, i.e.,*

$$\lim_{n\to\infty} H_n(t\sqrt{n}) = \exp\left(-\frac{t^2}{2}\right). \tag{5.17}$$

Proof. For the investigation of the behaviour of the kernel function when $n \to \infty$, we use the Taylor expansion of $H_n(q)$ centred at $q_0 = 0$

$$H_n(q) = \sum_{k=0}^{\infty} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2}+k)k!}\left(-\frac{q^2}{4}\right)^k, \tag{5.18}$$

and by using the substitution $q = t\sqrt{n}$, we can transform it into

$$H_n(t\sqrt{n}) = \sum_{k=0}^{\infty} \frac{1}{k!}\left(-\frac{t^2}{2}\right)^k \frac{\Gamma(\frac{n}{2})n^k}{\Gamma(\frac{n}{2}+k)2^k}. \tag{5.19}$$

For every $k \in \mathbb{N}$, it holds that

$$\lim_{n\to\infty} \frac{\Gamma(\frac{n}{2})n^k}{\Gamma(\frac{n}{2}+k)2^k} = 1, \tag{5.20}$$

which is a fundamental property of Gamma function. Therefore, the limit case of the kernel function equals

$$\lim_{n\to\infty} H_n(t\sqrt{n}) = \exp\left(-\frac{t^2}{2}\right). \tag{5.21}$$

□

For simplicity, we will use the following notation in the subsequent sections:

$$H_\infty(q) = \exp\left(-\frac{q^2}{2}\right). \tag{5.22}$$

This type of Gaussian kernel has been widely applied in the Parzen density estimates [58, 59, 60] of a probability density function, but with another meaning and motivation.

To sum up, altogether we have received two types of kernel function for rotational spectrum:

- $H_n(q)$ represents the rotation in n-dimensional space,
- $H_\infty(q)$ represents rotation in infinite-dimensional space.

The scaled kernel functions $H_n(q)$ are displayed in Fig. 5.2.

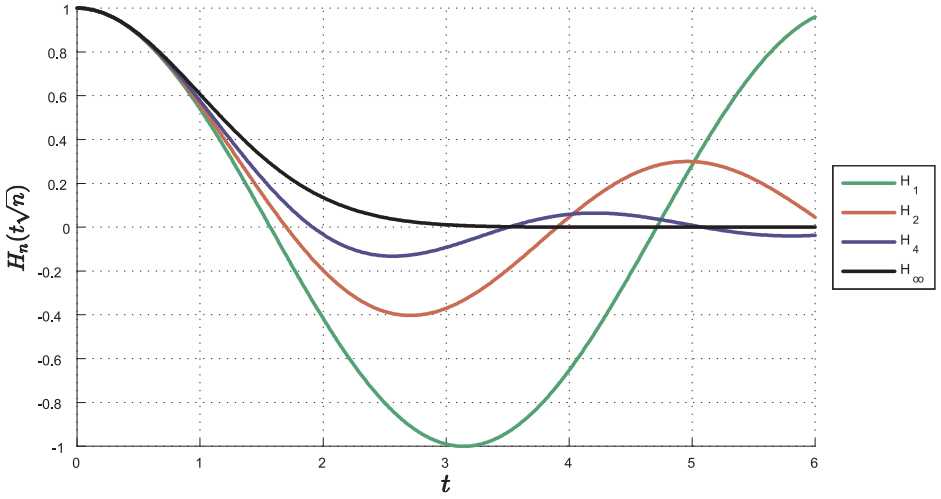


Figure 5.2: Scaled kernel functions $H_n(q)$ for rotational spectrum.

5.2 Expression using Characteristic Function

Analytical expression of rotational spectrum of set $\mathcal{F} \subset \mathbb{R}^n$

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})} H_n\left(\Omega\|\vec{x}-\vec{y}\|\right), \tag{5.23}$$

can be performed based on the probability density function (pdf) $f(\vec{x})$ of $\vec{x} \in \mathcal{F}$. Using the distribution we can determine the pdf of $\vec{z} = \vec{x} - \vec{y}$ where both $\vec{x}$ and $\vec{y}$ share the same pdf f. The pdf $g(\vec{z})$ of a point $\vec{z}$ is given by

$$g(\vec{z}) = (f \star f)(\vec{z}), \tag{5.24}$$

where $\star$ denotes the operator of correlation defined as

$$(f \star f)(\vec{z}) = \int_{\mathbb{R}^n} f^*\left(\vec{t}\right) f(\vec{z}+\vec{t})\,\mathrm{d}\vec{t} \tag{5.25}$$

and $*$ is the complex conjugate. The rotational spectrum can be expressed as

$$\begin{aligned} S(\Omega) &= \mathop{\mathrm{E}}_{\vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})} H_n\left(\Omega\|\vec{x}-\vec{y}\|\right) && (5.26)\\ &= \mathop{\mathrm{E}}_{\substack{\vec{z}=\vec{x}-\vec{y}\\ \vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})}} H_n\left(\Omega\|\vec{z}\|\right) && (5.27)\\ &= \int_{\mathbb{R}^n} g(\vec{z}) H_n(\Omega\|\vec{z}\|). && (5.28) \end{aligned}$$

We define the characteristic function of a set $\mathcal{G} \subset \mathbb{R}^n$ as

$$\psi(\vec{t}) = \mathop{\mathrm{E}}_{\vec{z}\sim \mathrm{U}(\mathcal{G})} \exp(\mathrm{i}\vec{t}\cdot\vec{z}) = \int_{\mathbb{R}^n} g(\vec{z}) \exp(\mathrm{i}\vec{t}\cdot\vec{z})\mathrm{d}\vec{z}. \tag{5.29}$$

Supposing that $g(\vec{z})$ is pdf and $\psi(\vec{t})$ is a set characteristic function, we can apply the inverse Fourier transform to function $g(\vec{t})$, yielding

$$\mathcal{F}_n^{-1}[g](\vec{t}) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} g(\vec{z})\exp(\mathrm{i}\vec{t}\cdot\vec{z})\mathrm{d}\vec{z} = \frac{1}{(2\pi)^n}\psi(\vec{t}) \tag{5.30}$$

therefore the distribution $g(\vec{z})$ can be expressed as

$$g(\vec{z}) = \frac{1}{(2\pi)^n}\mathcal{F}_n[\psi(\vec{t})](\vec{z}). \tag{5.31}$$

The rotational spectrum equals

$$\begin{aligned} S(\Omega) &= \int_{\mathbb{R}^n} g(\vec{z})H_n(\Omega\|\vec{z}\|)\mathrm{d}\vec{z} && (5.32)\\ &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^n}\mathcal{F}_n[\psi(\vec{t})]H_n(\Omega\|\vec{z}\|)\mathrm{d}\vec{z} && (5.33)\\ &= \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(\vec{t})\exp(-\mathrm{i}\vec{t}\cdot\vec{z})H_n(\Omega\|\vec{z}\|)\mathrm{d}\vec{z}\mathrm{d}\vec{t} && (5.34)\\ &= \int_{\mathbb{R}^n} \psi(\vec{t})\underbrace{\left(\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\exp(-\mathrm{i}\vec{t}\cdot\vec{z})H_n(\Omega\|\vec{z}\|)\mathrm{d}\vec{z}\right)}_{\Psi(\Omega,\vec{t})}\mathrm{d}\vec{t}, && (5.35)\end{aligned}$$

therefore

$$S(\Omega) = \int_{\mathbb{R}^n}\psi(t)\Psi(\Omega,\vec{t})\mathrm{d}\vec{t}, \tag{5.36}$$

where

$$\Psi(\Omega,\vec{t}) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\exp(-\mathrm{i}\vec{t}\cdot\vec{z})H_n(\Omega\|\vec{z}\|)\mathrm{d}\vec{z}. \tag{5.37}$$

Taking substitution $\vec{x} = \Omega\vec{z}$ we can rewrite into

$$\Psi(\Omega,\vec{t}) = \frac{1}{(2\pi\Omega)^n}\int_{\mathbb{R}^n}\exp\left(-\frac{\mathrm{i}\vec{t}\cdot\vec{x}}{\Omega}\right)H_n(\|\vec{x}\|)\mathrm{d}\vec{x}, \tag{5.38}$$

which can be further expressed using Fourier transform as

$$\Psi(\Omega,\vec{t}) = \frac{1}{(2\pi\Omega)^n}\mathcal{F}_n[H_n(\|\vec{x}\|)]\left(\frac{\vec{t}}{\Omega}\right) \tag{5.39}$$

Using inverse subtitution $\vec{t} = \Omega\vec{x}$ we can retrieve the equivalent expression for rotational spectrum as

$$S(\Omega) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\mathcal{F}_n[H_n(\|\vec{x}\|)](\vec{x})\,\psi(\Omega\vec{x})\mathrm{d}\vec{x}. \tag{5.40}$$

Since $H_n(\|\vec{x}\|)$ is a radial function, we can apply radial Fourier transform using Bochner formula [61] for $r = \|\vec{x}\|$ as

$$G(\xi) = \frac{\xi}{(2\pi\xi)^{\frac{n}{2}}}\int_0^\infty r^{\frac{n}{2}}H_n(r)J_{\frac{n-2}{2}}(\xi r)\mathrm{d}r, \tag{5.41}$$

yielding

$$S(\Omega) = \int_{\mathbb{R}^n} G(\|\vec{x}\|)\psi(\Omega\vec{x})\mathrm{d}\vec{x}. \tag{5.42}$$

If $G(\xi)$ fulfils the properties of a pdf of $\vec{x}$, we can express rotational spectrum as

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x}\sim G} \psi(\Omega\vec{x}). \tag{5.43}$$

However, function G does not necessarily have to be a pdf. For specific kernels H_n we will express the function explicitly to confirm the hypothesis.

We will be working with two kinds of kernel functions:

- $H_n(r)$ as the kernel function used for rotation,
- $G_n(\xi)$ its Fourier transform.

The function $G(\xi)$ is derived in case infinite-dimensional rotation where

$$H_\infty(r) = \exp\left(-\frac{r^2}{2}\right). \tag{5.44}$$

The integral contained in the radial Fourier transform equals

$$\int_0^\infty r^{\frac{n}{2}} H_\infty(r) J_{\frac{n-2}{2}}(\xi r)\mathrm{d}r = \int_0^\infty r^{\frac{n}{2}} \exp\left(-\frac{r^2}{2}\right) J_{\frac{n-2}{2}}(\xi r)\mathrm{d}r = \exp\left(-\frac{\xi^2}{2}\right)\xi^{\frac{n-2}{2}} \tag{5.45}$$

therefore, the function $G_\infty(\xi)$ equals

$$G_\infty(\xi) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{\xi^2}{2}\right). \tag{5.46}$$

This corresponds to the pdf of n-dimensional normal distribution with zero mean and unit covariance matrix. Thus, the infinite-dimensional rotation spectrum can be expressed as

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x}\sim\mathcal{N}(\vec{0},I)} \psi(\Omega\vec{x}) \tag{5.47}$$

The advantage of this expression is the possibility to use it for simulations. The rotational spectrum can be estimated using Monte Carlo approach using the known characteristic function.

As a next step, the $G(\xi)$ is simplified for finite dimensional rotation, therefore for any kernel function H_n and $n \in \mathbb{N}$. In this case, the kernel function equals

$$H_n(r) = \frac{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}{r^{\frac{n-2}{2}}} J_{\frac{n-2}{2}}(r). \tag{5.48}$$

Substituting the function into the formula for $G_n(\xi)$ and after simplification, we obtain

$$G_n(\xi) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\xi^{\frac{n-2}{2}}} \int_0^\infty r J_{\frac{n-2}{2}}(r) J_{\frac{n-2}{2}}(\xi r)\mathrm{d}r, \tag{5.49}$$

where the integral on the right hand side equals

$$\int_0^\infty r J_{\frac{n-2}{2}}(r) J_{\frac{n-2}{2}}(\xi r) \mathrm{d}r = \delta(\xi - 1), \tag{5.50}$$

where δ is Dirac function. Therefore $G_n(\xi)$ is degenerated pdf, corresponding to the uniform distribution of points on the surface of $(n-1)$ dimensional ball, which we denote $\mathcal{S}_{n-1}$. Finite-dimensional rotational spectrum can be subsequently expressed as

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x} \sim \mathrm{U}(\mathcal{S}_{n-1})} \psi(\Omega \vec{x}). \tag{5.51}$$

As easy to prove, in case of kernel function $H_0(r)$, the $G(\xi)$ does not satisfy the properties of a pdf, therefore the characteristic function approach to rotational spectrum estimation is not possible.

5.3 Limit Behaviour of Point Set Spectrum

In this section, we discuss the relationship between the rotational spectrum for the limit kernel H_∞ and the correlation dimension. The correlation sum is a cumulative distribution function of the distances between the points in a fractal set; therefore, the rotational spectrum can be written as a Stieltjes integral:

$$S(\Omega) = \int_0^\infty H_\infty(\Omega r) \mathrm{d}C(r) = \int_0^\infty \exp\left(-\frac{\Omega^2 r^2}{2}\right) \mathrm{d}C(r). \tag{5.52}$$

After the application of the integration by parts, we can obtain

$$S(\Omega) = \int_0^\infty \Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right) C(r) \Omega \mathrm{d}r, \tag{5.53}$$

and by substituting $\xi = \Omega r$, we get the integral formula for the rotational spectrum:

$$S(\Omega) = \int_0^\infty \xi \cdot C\left(\frac{\xi}{\Omega}\right) \exp\left(-\frac{\xi^2}{2}\right) \mathrm{d}\xi. \tag{5.54}$$

Theorem 3. *Let $\mathcal{F} \subset \mathbb{R}^n$ be a Lebesgue measurable set with the rotational spectrum*

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y} \sim U(\mathcal{F})} H_\infty(\Omega \left\| \vec{x} - \vec{y} \right\|), \tag{5.55}$$

and let us assume that correlation dimension D_2 exists. Then, it holds that

$$\lim_{\Omega \to \infty} \frac{\ln S(\Omega)}{\ln \Omega} = -D_2. \tag{5.56}$$

Proof. To prove this, let us suppose that $\delta < 1$ and that, at first, $r < \delta$. Assuming the existence of correlation dimension, we have $\forall \epsilon > 0 \ \exists \delta > 0$

$$0 < r < \delta \Rightarrow \left| \frac{\ln C(r)}{\ln r} - D_2 \right| < \epsilon,$$

therefore

$$r^{D_2+\epsilon} < C(r) < r^{D_2-\epsilon}. \tag{5.57}$$

However, for $r \geqslant \delta$, we have

$$\delta^{D_2+\epsilon} < C(r) \leqslant 1. \tag{5.58}$$

Now, we can estimate the lower and the upper boundary for the spectrum

$$S(\Omega) = \mathop{\mathrm{E}}_{\vec{x},\vec{y}\sim U(\mathcal{F})} \exp\left(-\Omega^2 \frac{\|\vec{x}-\vec{y}\|^2}{2}\right) = \int_0^\infty C(r)\Omega^2 r \exp\left(-\Omega^2 \frac{r^2}{2}\right) \mathrm{d}r \tag{5.59}$$

as

$$I_{\mathrm{L}}(\Omega) < S(\Omega) < I_U(\Omega). \tag{5.60}$$

We can rewrite I_U as

$$I_U(\Omega) = \int_0^\delta r^{D_2-\epsilon}\Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right) \mathrm{d}r + \int_\delta^\infty \Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right) \mathrm{d}r, \tag{5.61}$$

and after the substitution $t = \Omega^2 r^2/2$, we get

$$I_U(\Omega) = \Omega^{\epsilon-D_2} \cdot 2^{\frac{D_2-\epsilon}{2}} \cdot \int_0^{\Omega^2\delta^2/2} t^{\frac{D_2-\epsilon}{2}} \exp(-t)\mathrm{d}t + \exp\left(\frac{-\Omega^2\delta^2}{2}\right). \tag{5.62}$$

Therefore, the upper bound I_U can be expressed as

$$I_U(\Omega) = \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2-\epsilon} \cdot \frac{D_2-\epsilon}{2} \cdot \Gamma_{\mathrm{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2-\epsilon}{2}\right), \tag{5.63}$$

where Γ_{inc} is an incomplete Gamma function. It is possible to do an estimation from above as

$$I_U < \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2-\epsilon} \cdot \frac{D_2-\epsilon}{2} \cdot \Gamma\left(\frac{D_2-\epsilon}{2}\right). \tag{5.64}$$

The lower bound I_L is rewritten as

$$I_L(\Omega) = \int_0^\delta r^{D_2+\epsilon}\Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right) \mathrm{d}r + \int_\delta^\infty r^{D_2-\epsilon}\Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right) \mathrm{d}r \tag{5.65}$$

and can be estimated as

$$I_L(\Omega) > \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2+\epsilon} \cdot \Gamma_{\mathrm{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2+\epsilon}{2}+1\right). \tag{5.66}$$

Altogether, we receive the upper and the lower boundary for the logarithm of the rotational spectrum

$$(D_2-\epsilon)\left(\frac{1}{2}\ln 2 - \ln\Omega\right) + L_1(\Omega) > \ln S(\Omega) > (D_2+\epsilon)\left(\frac{1}{2}\ln 2 - \ln\Omega\right) + L_2(\Omega) \tag{5.67}$$

and after the rearrangement

$$-\epsilon + \frac{L_2(\Omega) + \frac{D_2+\epsilon}{2}\ln 2}{\ln\Omega} < \frac{S(\Omega)}{\ln\Omega} + D_2 < \epsilon + \frac{L_1(\Omega) + \frac{D_2-\epsilon}{2}\ln 2}{\ln\Omega} \tag{5.68}$$

for the functions

$$L_1(\Omega) = \ln \Gamma\left(\frac{D_2 - \epsilon}{2}\right) + \ln \frac{D_2 - \epsilon}{2} \tag{5.69}$$

and

$$L_2(\Omega) = \ln \Gamma_{\text{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2 + \epsilon}{2} + 1\right). \tag{5.70}$$

It holds that both L_1 and L_2 are constrained functions of Ω. Therefore, Ω_0 exists, which guarantees that, for any $\Omega > \Omega_0 > 1$, it is valid that

$$\left|\frac{\ln S(\Omega)}{\ln \Omega} + D_2\right| < 2\epsilon = \epsilon^*, \tag{5.71}$$

which completes the proof. □

The Lebesgue measurability of the investigated set is an important prerequisite because it ensures the capability to perform a uniform sampling. As a general remark, we could consider another kernel function instead of H_∞. For any non-increasing function $\Phi : \mathbb{R}_0^+ \mapsto [0; 1]$ satisfying $\Phi(0) = 1$ and $\Phi(\infty) = 0$, and whose first derivative $\Phi'(\xi)$ exists for any $\xi > 0$, we consider the rotational spectrum in a more general form as

$$S(\Omega) = \underset{\vec{x},\vec{y}\sim U(\mathcal{F})}{\mathrm{E}} \Phi\left(\Omega\|\vec{x} - \vec{y}\|\right). \tag{5.72}$$

The Ψ function is defined as

$$\Psi(\alpha) = -\int_0^\infty \xi^\alpha \Phi'(\xi))\mathrm{d}\xi, \tag{5.73}$$

and the existence of limit (5.56) is guaranteed only if both $\Psi(D_2 + \epsilon)$ and $\Psi(D_2 - \epsilon)$ are finite for arbitrary $\epsilon \in (0; \epsilon_0)$. Another example of a kernel function could be the generalised exponential kernel

$$\Phi_1(\xi) = \exp\left(-\frac{\xi^\beta}{\beta}\right) \tag{5.74}$$

for $\beta > 0$ or the inverse polynomial kernel

$$\Phi_2(\xi) = \frac{1}{P(\xi)}, \tag{5.75}$$

where $P(\xi)$ represents a polynomial of order $M > D_2 + 1$.

5.4 Estimation of D_2

The simulation of the rotational spectrum is based on generating point pairs using a Monte Carlo approach. The points are independently and uniformly sampled from the analysed set $\mathcal{F}$. With $M \in \mathbb{N}$ fixed and $\vec{x}_i, \vec{y}_i \sim U(\mathcal{F})$, the rotational spectrum is estimated as

$$\widehat{S}(\Omega) = \frac{1}{M}\sum_{j=1}^{M} H_n\left(\Omega\|\vec{x}_j - \vec{y}_j\|\right) \tag{5.76}$$

including the variance estimate

$$\widehat{\text{var}\, S}(\Omega) = \frac{1}{M-1} \sum_{j=1}^{M} \left(H_n \left(\Omega \| \vec{x}_j - \vec{y}_j \| \right) - \widehat{S}(\Omega) \right)^2 \tag{5.77}$$

To take advantage of the linear dependence between the logarithm of the rotational spectrum and the logarithm of the distance, we can reasonably consider the model

$$\ln S(\Omega) = A - D_2 \cdot \ln\, \Omega + \epsilon. \tag{5.78}$$

The estimation of parameter D_2 is based on the maximum likelihood method using L_p regression with a minimisation criterion

$$CRIT = \sum_{k=1}^{N} |y_k - f(x_k, \vec{a})|^p \tag{5.79}$$

for $p > 1$ and a general model formulated as $y = f(x_k, \vec{a})$. In our case, the minimisation criterion satisfies

$$CRIT^* = \sum_{k=1}^{N} \left| \ln\, \widehat{S}(\Omega_k) - A + D_2 \ln\, \Omega_k \right|^p. \tag{5.80}$$

The algorithm is based on the capability to generate point pairs uniformly from a fractal set and can be formulated as follows:

- The parameter M is chosen arbitrarily, but is large enough (e.g. $M = 10^5$). This parameter represents the number of Monte Carlo simulations, which is equal to the number of point pairs from the fractal set used for the estimation.
- The values of Ω, in which the calculation is performed, are determined. For the simulation, it is recommended to choose regular sampling from the interval, where the rotational spectrum is expected to have a linear characteristic.
- The calculation of the rotational spectrum is performed at points $\Omega_1, \Omega_2 \ldots, \Omega_N$, according to equation (5.76).
- With the values of Ω_i and the respective $\widehat{S}(\Omega_i)$ estimates, it is possible to perform minimisation using equation (5.80) with a maximum likelihood method.
- The resulting parameter D_2 represents the estimate of the correlation dimension.

The algorithm's complexity is similar as in the case of traditional correlation sum, since it allows to generate point pairs and calculate the rotational spectrum $\widehat{S}(\Omega)$ directly for any Ω. We will prove in the experimental part that the dependency in the spectral domain is smoother than in the time domain.

Chapter 6

Modified Renyi Dimension Estimation

The chapter describes an alternative method for estimation of Renyi dimension D_α. The traditional approach of determination of Renyi dimension is based on Renyi entropy estimate, which is biased in general. The second and more general problem is the method of sampling of the point set. Here we focus only on the Lebesgue measurable sets where uniform sampling is defined. When these conditions are not guaranteed, such as when the geometric structure of the set is inhomogeneous, we can only test the hypothesis of unbiasedness for the given theoretical value D_α i.e. $H_0 : \widehat{D_\alpha} = D_\alpha$.

6.1 Renyi Entropy Estimation via Parzen Approach

This section utilizes the Parzen estimate for the derivation of the density function of elements of the Lebesgue measurable set $\mathcal{F} \subset \mathbb{R}^n$. Supposing the existence of n-dimensional distribution function ϕ of points $\vec{x} \in \mathcal{F}$ i.e. $\vec{x} \sim \phi$, it is possible to define a sample of points

$$\Phi = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_M}\} \subset \mathcal{F} \tag{6.1}$$

that are uniformly generated from $\mathcal{F}$, i.e. $\vec{x_k} \sim \mathrm{U}(\mathcal{F}) \equiv \phi$. For any point $\vec{x} \in \mathbb{R}^n$, we define its ϵ-neighbourhood, i.e. a ball with radius ϵ as

$$\mathcal{B}(\vec{x}, \epsilon) = \{\vec{y} \in \mathbb{R}^n : \|\vec{y} - \vec{x}\|_2 \leqslant \epsilon\} \tag{6.2}$$

for any $\epsilon > 0$. The volume V_* of the ball can be expressed as

$$V_* = V_n \cdot \epsilon^n \tag{6.3}$$

where V_n is the volume of an n-dimensional unit ball. The density estimate will be based on the elementary distribution

$$f_0(\vec{x}, \epsilon) = \frac{\mathrm{I}(\|\vec{x}\|_2 \leqslant \epsilon)}{V_*} \tag{6.4}$$

using the indicator function I(...). We can use Parzen's [62] formula

$$f(\vec{x}, \Phi, \epsilon) = \frac{1}{M} \sum_{k=1}^{M} f_0(\vec{x} - \vec{x_k}, \epsilon) \tag{6.5}$$

to obtain a consistent estimate of ϕ. However, we will apply (6.5) to a discontinuous distribution on $\mathcal{F}$ to obtain new formulas for the Renyi dimension estimation. The probability density estimate (6.5) is visualised in Fig. 6.1 in grayscale. The white area represents the regions where this function equals zero and the darker areas depict the intersection of several balls centred at points from the set sample Φ.

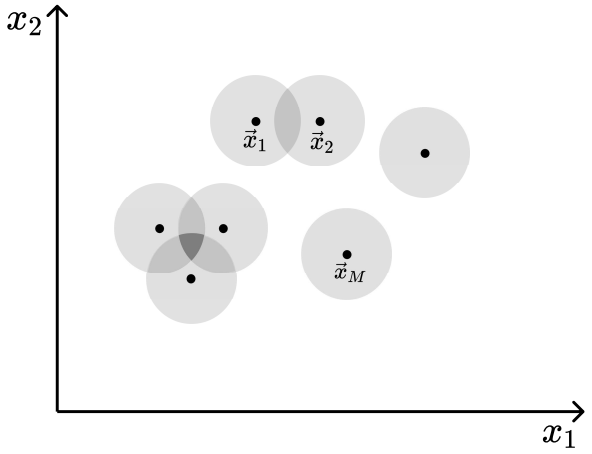


Figure 6.1: Density of Parzen estimate.

6.2 Limit Behaviour of Entropy Slope

Our novel estimate of the Renyi entropy is based on the differential entropy

$$H_\alpha = \frac{1}{1-\alpha} \ln \int_{\vec{x} \in \mathbb{R}^n} f^\alpha(\vec{x}) \mathrm{d}\vec{x} \tag{6.6}$$

for $\alpha \in \mathbb{R}_0^+ \setminus \{1\}$ and the Parzen estimate $f(\vec{x})$ that is scaled by $\epsilon > 0$. To avoid negative entropy values, we define the modified Renyi entropy as

$$H_\alpha^*(\Phi, \epsilon) = \frac{\ln J(\Phi, \alpha, \epsilon) - \ln J_0(\alpha, \epsilon)}{1-\alpha}. \tag{6.7}$$

for $\alpha \geqslant 0$ and $\alpha \neq 1$ where

$$J(\Phi, \alpha, \epsilon) = \int_{\vec{x} \in \mathbb{R}^n} f^\alpha(\vec{x}, \Phi, \epsilon) \mathrm{d}\vec{x} \tag{6.8}$$

and

$$J_0(\alpha, \epsilon) = \int_{\vec{x} \in \mathbb{R}^n} f_0^{\alpha}(\vec{x}, \epsilon) \mathrm{d}\vec{x} = V_*^{1-\alpha}. \tag{6.9}$$

Using the expected value of $v(\vec{x})$ for $\vec{x} \sim g$ as

$$\underset{\vec{x} \sim g}{\mathrm{E}}\, v(\vec{x}) = \int_{\vec{x} \in \mathbb{R}^n} v(\vec{x}) g(\vec{x}) \mathrm{d}\vec{x}, \tag{6.10}$$

the first term can be simplified as

$$J = J(\Phi, \alpha, \epsilon) = \int_{\vec{x} \in \mathbb{R}^n} f^{\alpha-1}(\vec{x}, \Phi, \epsilon) \cdot f(\vec{x}, \Phi, \epsilon) \mathrm{d}\vec{x} = \underset{\vec{x} \sim f}{\mathrm{E}} \quad f^{\alpha-1}(\vec{x}, \Phi, \epsilon). \tag{6.11}$$

For the sample $\Phi = \{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_M\}$ we define the degeneracy of $\vec{x} \in \mathbb{R}^n$ as

$$G(\vec{x}, \Phi, \epsilon) = \sum_{k=1}^{M} \mathrm{I}(\|\vec{x} - \vec{x_k}\|_2 \leqslant \epsilon) \tag{6.12}$$

holding that $G(\vec{x}, \Phi, \epsilon) \in \{0, \ldots, M\}$. Recall that the probability density function $f(\vec{x})$ is

$$f(\vec{x}, \Phi, \epsilon) = \frac{1}{M \cdot V_*} \sum_{k=1}^{M} \mathrm{I}(\|\vec{x} - \vec{x_k}\|_2 \leqslant \epsilon) = \frac{G(\vec{x}, \Phi, \epsilon)}{M \cdot V_*}. \tag{6.13}$$

Therefore

$$J = \underset{\vec{x} \sim f}{\mathrm{E}} \quad \left(\frac{G(\vec{x}, \Phi, \epsilon)}{M \cdot V_*} \right)^{\alpha-1} = M^{1-\alpha} V_*^{1-\alpha} \underset{\vec{x} \sim f}{\mathrm{E}} G^{\alpha-1}(\vec{x}, \Phi, \epsilon) \tag{6.14}$$

and subsequently also the modified Renyi entropy is

$$H_\alpha^*(\Phi, \epsilon) = \frac{\ln J - \ln J_0}{1 - \alpha} = \tag{6.15}$$

$$= \frac{(1-\alpha) \ln M + (1-\alpha) \ln V_* + \ln \mathrm{E}\, G^{\alpha-1}(\vec{x}, \Phi, \epsilon)}{1 - \alpha} - \tag{6.16}$$

$$- \frac{(1-\alpha) \ln V_*}{1 - \alpha}. \tag{6.17}$$

The resulting modified entropy equals

$$H_\alpha^*(\Phi, \epsilon) = \ln M + \frac{\ln \mathrm{E}\, G^{\alpha-1}(\vec{x}, \Phi, \epsilon)}{1 - \alpha} \tag{6.18}$$

for $\alpha > 0$ and $\alpha \neq 1$.

The degeneracy of $\vec{x} \in \mathbb{R}^n$ will be further denoted as $G(\vec{x})$ instead of $G(\vec{x}, \Phi, \epsilon)$ for simplicity. When $\vec{x} \in \mathbb{R}^n$, the degeneracy $G(\vec{x}) \in \{0, \ldots, M\}$, but for $\vec{x} \in \Phi$, the degeneracy fulfils $G(\vec{x}) \in \{1, \ldots, M\}$. The modified Renyi entropy follows $0 \leqslant H_\alpha^* \leqslant \ln M$. This entropy is a translational and rotational invariant, as it is easy to prove from (6.12) and (6.18). For the particular cases of α, one can derive the

- Modified Hartley entropy for $\alpha = 0$ as

$$H_0^* = \ln M + \ln \mathrm{E} G^{-1}(\vec{x}), \tag{6.19}$$

- Modified Shannon entropy as a limit for $\alpha \to 1$ i. e.

$$H_1^* = \lim_{\alpha \to 1} H_\alpha^* = \ln M - \mathrm{E}\ \ln G(\vec{x}), \tag{6.20}$$

- Modified collision entropy for $\alpha = 2$ as

$$H_2^* = \ln M - \ln \mathrm{E}\, G(\vec{x}), \tag{6.21}$$

- Modified minimum entropy as a limit for $\alpha \to \infty$ as

$$H_\infty^* = \lim_{\alpha \to +\infty} H_\alpha^* = \ln M - \ln \max G(\vec{x}). \tag{6.22}$$

where the expected values are over $\vec{x} \sim f$. The dependency of the modified Renyi entropies on the logarithm of ϵ is displayed in Fig. 6.2. If the derivative $\frac{\partial H^*}{\partial \alpha}$ exists, it is always less or equal to zero, as it is easy to prove.

Moreover, the modified Renyi entropy H_α^* can be used for an alternative definition of the dimension as

$$D_\alpha^* = \lim_{\epsilon \to 0^+} \frac{H_\alpha^*(\epsilon)}{-\ln \epsilon} \tag{6.23}$$

for a given $\mathcal{F}$ as an analogy to formula (1.14).

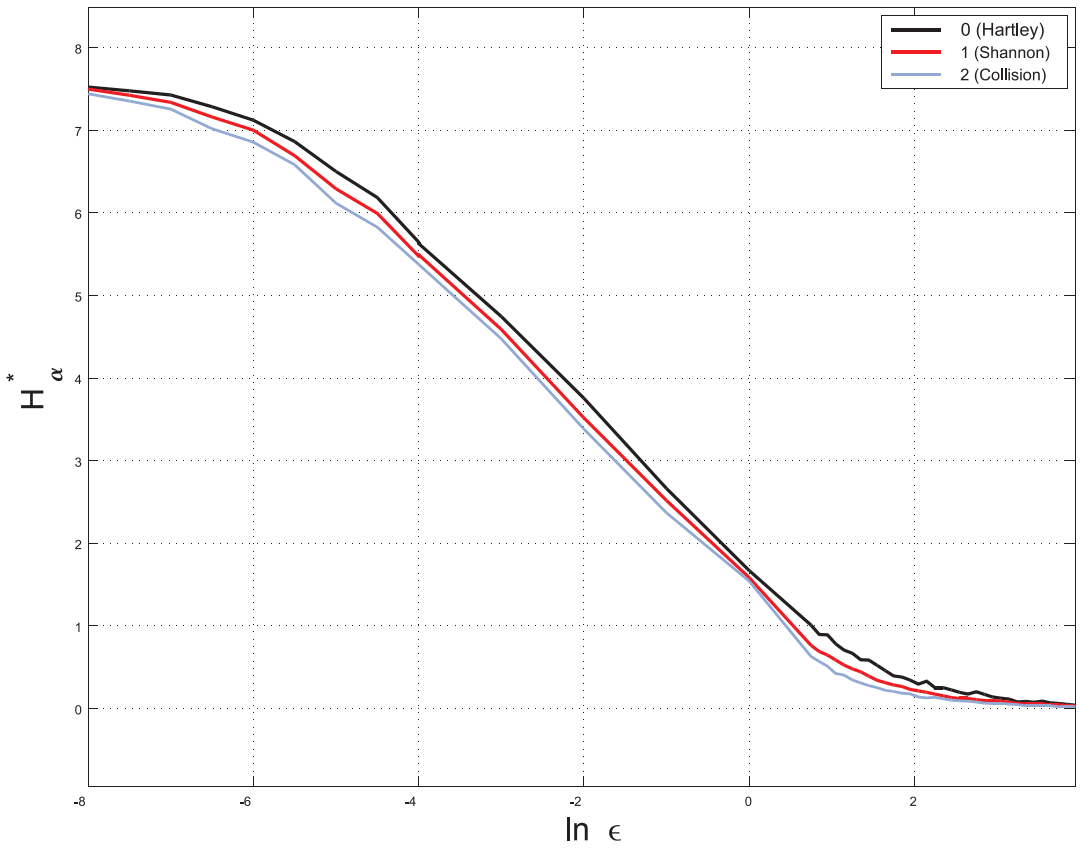


Figure 6.2: Modified Renyi entropies for Cantor set.

6.3 Relationship to Capacity and Correlation Dimensions

The capacity (D_0) and correlation (D_2) dimensions are defined for any Lebesgue measurable set $\mathcal{F}$. The only possibility how to compare D_α^* with D_α is to come back from the sample Φ to the original set $\mathcal{F}$. The sample Φ is a finite set with $D_\mathrm{H} = D_0 = D_2 = 0$, of course. We will study the particular cases of D_α^* for $\alpha = 0$ and $\alpha = 2$ in the case of the measurable set $\mathcal{F}$. As we see in the following theorem, the Renyi dimension is the characteristic that has an important relationship to both capacity and correlation dimensions.

Theorem 4. *Definition of D_0^* is equivalent to the capacity dimension D_0 for any measurable set $\mathcal{F}$.*

Proof. The capacity dimension can be defined [3] based on the Minkowski sausage as

$$D_0 = n - \lim_{\epsilon \to 0^+} \frac{\ln \mu(\mathcal{S})}{\ln \epsilon} \tag{6.24}$$

where

$$\mathcal{S} = \bigcup_{\vec{x} \in \mathcal{F}} \mathcal{B}(\vec{x}, \epsilon). \tag{6.25}$$

and $\mu(\mathcal{S}) = \int_{\vec{x} \in \mathcal{S}}$ is its volume. Supposing the existence of D_0, we can directly calculate

$$J_0 = \int_{\mathbb{R}^n} f_0^0(\vec{x}) \mathrm{d}\vec{x} = V_* = V_n \cdot \epsilon^n, \tag{6.26}$$

and also the density

$$f(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}) \tag{6.27}$$

and

$$f^0(\vec{x}) = \mathrm{I}\left(\mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}) > 0 \right) = \mathrm{I}\left(\bigvee_{\vec{y} \in \mathcal{F}} \|\vec{x} - \vec{y}\|_2 \leqslant \epsilon \right) = \mathrm{I}(\vec{x} \in \mathcal{S}). \tag{6.28}$$

Therefore, the function J can be expressed as

$$J = \int_{\mathbb{R}^n} f^0(\vec{x}) \mathrm{d}\vec{x} = \mu(\mathcal{S}). \tag{6.29}$$

The resulting modified Hartley entropy equals

$$H_0^*(\epsilon) = \ln \frac{\int_{\mathbb{R}^n} f^0(\vec{x}) \mathrm{d}\vec{x}}{\int_{\mathbb{R}^n} f_0^0(\vec{x}) \mathrm{d}\vec{x}} = \ln \mu(\mathcal{S}) - \ln V_* = \ln \mu(\mathcal{S}) - \ln V_n - n \ln \epsilon. \tag{6.30}$$

Now, it is clear that

$$D_0^* = \lim_{\epsilon \to 0^+} \frac{H_0^*(\epsilon)}{-\ln \epsilon} \tag{6.31}$$

$$= \lim_{\epsilon \to 0^+} \frac{\ln \mu(\mathcal{S}) - \ln V_n - n \ln \epsilon}{-\ln \epsilon} = \tag{6.32}$$

$$= n - \lim_{\epsilon \to 0^+} \frac{\ln \mu(\mathcal{S})}{\ln \epsilon} = D_0. \tag{6.33}$$

Therefore, D_0^* obtained from the modified Hartley entropy $H_0^*(\epsilon)$ is equivalent to the capacity dimension D_0 of the measurable set $\mathcal{F}$. □

Theorem 5. *Definition of D_2^* is equivalent to the correlation dimension D_2 for any measurable set $\mathcal{F}$.*

Proof. The correlation dimension of $\mathcal{F}$ is defined as

$$D_2 = \lim_{\epsilon \to 0^+} \frac{\ln C(\epsilon)}{\ln \epsilon} \tag{6.34}$$

where

$$C(\epsilon) = \mathop{\mathrm{E}}_{\vec{y},\vec{z} \sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y} - \vec{z}\|_2 \leqslant \epsilon) \tag{6.35}$$

is the correlation integral. Supposing the existence of D_2, recall that

$$f_0(\vec{x}) = \frac{\mathrm{I}(\|\vec{x}\|_2 \leqslant \epsilon)}{V_n \epsilon^n} \tag{6.36}$$

and by means of integrating the elementary distribution over the space we get

$$J_0 = \int_{\mathbb{R}^n} f_0^2(\vec{x}) \mathrm{d}\vec{x} = \frac{1}{V_n^2 \epsilon^{2n}} V_n \epsilon^n = V_n^{-1} \epsilon^{-n}. \tag{6.37}$$

In the finite case, we have

$$f(\vec{x}) = \frac{1}{m} \sum_{k=1}^{m} f_0(\vec{x} - \vec{x_k}), \tag{6.38}$$

which can be generalized to

$$f(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}). \tag{6.39}$$

Therefore,

$$f^2(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y},\vec{z} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}) \cdot f_0(\vec{x} - \vec{z}) \tag{6.40}$$

and

$$J = \int_{\mathbb{R}^n} f^2(\vec{x}) \mathrm{d}\vec{x} = \mathop{\mathrm{E}}_{\vec{y},\vec{z} \sim \mathrm{U}(\mathcal{F})} Z(\vec{y}, \vec{z}) \tag{6.41}$$

where

$$Z(\vec{y}, \vec{z}) = \int_{\mathbb{R}^n} f_0(\vec{x} - \vec{y}) \cdot f_0(\vec{x} - \vec{z}) \mathrm{d}\vec{x}. \tag{6.42}$$

Denoting the distance $d = \|\vec{y} - \vec{z}\|_2$, we can evaluate

$$Z(\vec{y}, \vec{z}) = V_n^{-2} \epsilon^{-2n} \int_{\mathbb{R}^n} \mathrm{I}(\|\vec{x} - \vec{y}\|_2 \leqslant \epsilon) \mathrm{I}(\|\vec{x} - \vec{z}\|_2 \leqslant \epsilon) \mathrm{d}\vec{x} = V_n^{-2} \epsilon^{-2n} W_n(d, \epsilon) \tag{6.43}$$

where $W_n(d, \epsilon)$ is the volume of two hyper-ball intersections in the case of the mutual center distance d and radii ϵ. Using n-dimensional analytical geometry, we obtain

$$W_n(d, \epsilon) = 2 \int_{d/2}^{\epsilon} V_{n-1} \left(\epsilon^2 - r^2\right)^{\frac{n-1}{2}} \mathrm{d}r = 2 V_{n-1} \epsilon^n \int_{d/2\epsilon}^{1} \left(1 - r^2\right)^{\frac{n-1}{2}} \mathrm{d}r \tag{6.44}$$

and after substitution $r = \cos\phi$, we get

$$W_n(d,\epsilon) = 2V_{n-1}\epsilon^n \int_0^{\arccos(d/2\epsilon)} \sin^n\phi \ \mathrm{d}\phi. \tag{6.45}$$

Moreover,

$$W_n(0,\epsilon) = 2V_{n-1}\epsilon^n \int_0^{\pi/2} \sin^n\phi \ \mathrm{d}\phi = V_n \cdot \epsilon^n \tag{6.46}$$

which is also the volume of the n-dimensional ball of radius ϵ. Therefore, we can express the Z function as

$$Z(\vec{y},\vec{z}) = V_n^{-1}\epsilon^{-n}\frac{\int_0^{\arccos(d/2\epsilon)} \sin^n\phi \ \mathrm{d}\phi}{\int_0^{\pi/2} \sin^n\phi \ \mathrm{d}\phi} \tag{6.47}$$

and the entropy is

$$H_2^*(\epsilon) = -\ln \mathop{\mathrm{E}}_{\vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{x}-\vec{y}\|_2,\epsilon) \tag{6.48}$$

where

$$S_n(d,\epsilon) = \frac{\int_0^{\arccos(d/2\epsilon)} \sin^n\phi \ \mathrm{d}\phi}{\int_0^{\pi/2} \sin^n(\phi)\mathrm{d}\phi} \tag{6.49}$$

for $0 \leqslant d < 2\epsilon$ and $S_n(d,\epsilon) = 0$ for $d \geqslant 2\epsilon$. Let

$$Q_n = \frac{\int_0^{\pi/3} \sin^n\phi \ \mathrm{d}\phi}{\int_0^{\pi/2} \sin^n\phi \ \mathrm{d}\phi} \in (0,1) \tag{6.50}$$

be the value of $S_n(\epsilon,\epsilon)$. When $0 \leqslant d \leqslant \epsilon$, we can estimate the ratio as

$$Q_n \leqslant S_n(d,\epsilon) \leqslant 1. \tag{6.51}$$

For $\epsilon < d \leqslant 2\epsilon$, we have $0 \leqslant S_n(d,\epsilon) < Q_n$. Therefore, we can underestimate

$$S_n(d,\epsilon) \geqslant \mathrm{I}(d \leqslant \epsilon)\cdot Q_n \tag{6.52}$$

and an adequate upper estimate is

$$S_n(d,\epsilon) \leqslant \mathrm{I}(d \leqslant \epsilon) + (\mathrm{I}(d \leqslant 2\epsilon) - \mathrm{I}(d \leqslant \epsilon))\cdot Q_n = \tag{6.53}$$

$$(1-Q_n)\cdot \mathrm{I}(d \leqslant \epsilon) + Q_n \cdot \mathrm{I}(d \leqslant 2\epsilon) \leqslant \tag{6.54}$$

$$\leqslant (1-Q_n)\cdot \mathrm{I}(d \leqslant 2\epsilon) + Q_n \cdot \mathrm{I}(d \leqslant 2\epsilon) = \mathrm{I}(d \leqslant 2\epsilon). \tag{6.55}$$

We can continue in the estimation to obtain

$$\mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y}-\vec{z}\|_2 \leqslant \epsilon)\cdot Q_n \leqslant \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{y}-\vec{z}\|_2,\epsilon) \tag{6.56}$$

$$\leqslant \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y}-\vec{z}\|_2 \leqslant 2\epsilon) \tag{6.57}$$

and therefore

$$L(\epsilon) \leqslant \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{y}-\vec{z}\|_2,\epsilon) \leqslant U(\epsilon) \tag{6.58}$$

where the lower bound equals

$$L(\epsilon) = Q_n \cdot C(\epsilon) \tag{6.59}$$

and the appropriate upper bound equals

$$U(\epsilon) = C(2\epsilon). \tag{6.60}$$

For all $0 < \epsilon < 1$, the following inequalities hold

$$\frac{\ln U(\epsilon)}{\ln \epsilon} \leqslant \frac{H_2^*(\epsilon)}{-\ln \epsilon} \leqslant \frac{\ln L(\epsilon)}{\ln \epsilon}. \tag{6.61}$$

We can calculate

$$\lim_{\epsilon \to 0^+} \frac{\ln L(\epsilon)}{\ln \epsilon} = \lim_{\epsilon \to 0^+} \left(\frac{\ln Q_n}{\ln \epsilon} + \frac{\ln C(\epsilon)}{\ln \epsilon} \right) = D_2 \tag{6.62}$$

and also for the upper bound

$$\lim_{\epsilon \to 0^+} \frac{\ln U(\epsilon)}{\ln \epsilon} = \lim_{\epsilon \to 0^+} \frac{\ln C(2\epsilon)}{\ln \epsilon} = \lim_{\epsilon \to 0^+} \frac{\ln C(\epsilon)}{\ln \frac{\epsilon}{2}} =$$

$$= \lim_{\epsilon \to 0^+} \frac{\ln C(\epsilon)}{\ln \epsilon} \cdot \lim_{\epsilon \to 0^+} \frac{\ln \epsilon}{\ln \epsilon - \ln 2} = D_2. \tag{6.63}$$

Therefore,

$$D_2^* = \lim_{\epsilon \to 0^+} \frac{H_2^*(\epsilon)}{-\ln \epsilon} = D_2. \tag{6.64}$$

□

We have proven that the alternative definition of Renyi dimension is equivalent to the definition of capacity and correlation dimensions. As a hypothesis that will be validated experimentally, we will examine whether this relationship holds for other $\alpha > 0$.

6.4 Estimation of D_α

Using the operator U of uniform sampling, the approximation of the Renyi entropy can be achieved via Monte Carlo technique in the following way:

1. At first, the sample index is generated uniformly $k \sim \mathrm{U}(\{1, \ldots, M\})$.
2. The point $\vec{x}$ is generated uniformly from the ϵ-ball centred at $\vec{x_k}$ as $\vec{x} \sim \mathrm{U}(\mathcal{B}(\vec{x_k}, \epsilon))$.
3. The subsequent degeneration is calculated using (6.12).

The entropy H_α^* is calculated as an average of the degenerations using (6.18), (6.20) or (6.22) depending on α. The first two steps generate $\vec{x} \sim f$, of course. Assuming the entropy estimate H_α^* fulfils $H_\alpha^* \propto \epsilon^{-D_\alpha^*}$ for small $\epsilon > 0$, we can use it for the estimation of D_α^* using the model

$$H_\alpha^*(\epsilon) = A - D_\alpha^* \ln \epsilon \tag{6.65}$$

for small ϵ and satisfying linear dependency H_α^* on $\ln \epsilon$.

Chapter 7

Revisited Zero-Crossing for Fractional Signal Analysis

This chapter summarizes the main contributions to the traditional zero-crossing method that serves to estimate the Hurst exponent of fBm sample. The novel method uses Bayesian approach, thus provides more information about the fractal relevance of investigated data and thanks to the segmentation employs more features calculated at the interval level.

7.1 Revisited Zero-Crossing Method

The number of zero-crossings (or level crossings) of Gaussian processes and its relevant fractal properties were published in several papers by Azais [50], Feuerverger [51] and Couerjolly [48]. The main result of these studies is the dependence of Hurst parameter on the probability of zero-crossing of random continuous nowhere differentiable process at fixed interval. Our aim is to improve the method by analysing the number of zero-crossings of fGn process and to deduce the formula in another way.

Suppose $x_k = x_k(H)$ is a sampling of fGn process with parameter H, zero mean and unit variance on fixed interval for $k = 1, ..., N+1$. The corresponding autocorrelation function ρ_k [2] equals

$$\rho_k = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right). \tag{7.1}$$

We denote the first autocorrelation coefficient by $\rho \equiv \rho_1 = 2^{2H-1} - 1$. Since it is a Gaussian process, the probability density function of two consecutive sample elements is

$$f(x_k, x_{k+1}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x_k^2 - 2\rho x_k x_{k+1} + x_{k+1}^2\right)\right), \tag{7.2}$$

therefore the probability of zero-crossing equals

$$p^* = \text{prob}\left(x_k x_{k+1} < 0\right) \tag{7.3}$$

that can be expressed as

$$p^* = \frac{1}{\pi\sqrt{1-\rho^2}} \int_0^{+\infty}\int_0^{+\infty} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2+2\rho xy+y^2\right)\right) \mathrm{d}x\mathrm{d}y. \tag{7.4}$$

Using the substitutions $\xi = (\rho x + y)(1-\rho^2)^{-\frac{1}{2}}$ and $\eta = x$ one obtains

$$p^* = \frac{1}{\pi} \iint_{\mathcal{O}^*} \exp\left(-\frac{\xi^2+\eta^2}{2}\right) \mathrm{d}\xi\mathrm{d}\eta \tag{7.5}$$

where $\mathcal{O}^* = \left\{\left(\rho x + y), x\sqrt{1-\rho^2}\right) \in \mathbb{R}^2 : x, y \in \mathbb{R}_0^+\right\}$. Substituting $\xi = r\cos\varphi$ and $\eta = r\sin\varphi$ the crossing probability can be expressed as

$$p^* = \frac{1}{\pi} \int_{\varphi_{\min}}^{\varphi_{\max}} \mathrm{d}\varphi \cdot \int_0^{+\infty} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r = \frac{\varphi_{\max} - \varphi_{\min}}{\pi}. \tag{7.6}$$

Due to the form of the area $\mathcal{O}^*$ we have $\varphi_{\min} = 0$ and $\varphi_{\max}$ equals

$$\varphi_{\max} = \begin{cases} \operatorname{atan}\frac{\sqrt{1-\rho^2}}{\rho} & \rho > 0 \\ \pi - \operatorname{atan}\frac{\sqrt{1-\rho^2}}{|\rho|} & \text{otherwise} \end{cases}. \tag{7.7}$$

Utilizing equations (7.6) and (7.7), it is possible to express the probability p^* as

$$p^* = \frac{1}{\pi}\operatorname{arccot}\frac{\rho}{\sqrt{1-\rho^2}} \tag{7.8}$$

and finally using the identity $\rho = 2^{2H-1} - 1$, the point estimate of Hurst parameter H equals

$$H = 1 + \log_2 \cos\frac{\pi p^*}{2}, \tag{7.9}$$

which is revisited form of well-known formula published by Couerjolly [48], originally in the form

$$H = \frac{1}{2}\left(1 + \log_2(1 + \epsilon|\cos(\pi \cdot S_N)|)\right), \tag{7.10}$$

where the relative number of zero crossing S_N with respect to the sample length N and $\epsilon = \operatorname{sgn}\left(H - \frac{1}{2}\right)$ are assumed to be known.

7.2 Bayesian Approach

We already know how to obtain point estimate of H with given fGn sample of length $N+1$ – if the number of zero-crossings in the sample equals $Z > 0$, we can approximately express H as

$$H \approx 1 + \log_2 \cos\frac{\pi Z}{2N}. \tag{7.11}$$

Unfortunately, from formula (7.11) we do not know the distribution of Hurst exponent H neither the distribution of zero-crossings p^*. Suppose, that the probability p^* is known for investigated sample. Then the number of zero-crossings Z comes from binomial distribution $Z \sim \mathrm{Bi}(N, p^*)$ with probability density function

$$f(Z|p^*) = \binom{N}{Z}(p^*)^Z \cdot (1-p^*)^{N-Z}. \tag{7.12}$$

Using the Bayesian inverse rule [63], the posterior probability density equals

$$f(p^*|Z) = \frac{(p^*)^Z(1-p^*)^{N-Z}}{B(N+1, N-Z+1)}, \tag{7.13}$$

where $B(p, q)$ is the standard beta function. We can generalize the result by means of using Beta prior as

$$f_{\mathrm{PRIOR}}(p^*) = \frac{(p^*)^{-a}(1-p^*)^{-a}}{B(1-a, 1-a)}, \tag{7.14}$$

for parameter $a \in [0, 1)$. After the application of Bayesian rule, we get the posterior probability for p^* in the form

$$f_{\mathrm{POST}}(p^*|Z) = \frac{(p^*)^{Z-a}(1-p^*)^{N-Z-a}}{B(Z+1-a, N-Z+1-a)}, \tag{7.15}$$

therefore the probability p^* comes from Beta distribution $p^* \sim \mathrm{B}(Z+1-a, N-Z+1-a)$. Different estimates of p^* can be obtained for different choices of parameter a, however it is reasonable to consider only natural Bayesian approach with $a = 0$ or Jeffreys-Perks law [64],[65] with $a = 1/2$.

7.3 Signal Segmentation

To be able to capture changes in fractional character over time, it is necessary to divide the original sampling into disjoint segments and determine the statistical features in each interval separately. Function (7.15) describes the probability of zero-crossing in the fGn sample in general – considering the segmentation into L disjoint segments, each with $N+1$ elements, the probability $f_{\mathrm{L}}(p)$ of zero-crossing of the whole signal sample with $L \cdot (N+1)$ elements equals

$$f_{\mathrm{L}}(p) = \frac{1}{L}\sum_{k=1}^{L} \frac{p^{Z_k-a}(1-p)^{N-a-Z_k}}{B(Z_k+1-a, N-Z_k+1-a)}, \tag{7.16}$$

where Z_k is the number of zero-crossings in k-th segment. Based on the shape of the aggregate probability density $f_{\mathrm{L}}(p)$ one can calculate the mean value of Hurst parameter as

$$\mathrm{E}(H) = 1 + \int_0^1 f_L(p) \log_2 \cos\frac{p\pi}{2} \mathrm{d}p \tag{7.17}$$

and variance

$$\sigma^2 = \int_0^1 \left(1 + \log_2 \cos \frac{p\pi}{2}\right)^2 f_L(p)\mathrm{d}p - (\mathrm{E}\ H)^2. \tag{7.18}$$

with fixed number of signal divisions L.

As easy to prove, fGn of length $N \in \mathbb{N}$ has the same mean value and variance of H estimate for any nonoverlaping segmentation of constant segment length M, where $M < N$. But in the case of real signal, the fGn assumption with constant Hurst parameter is not always fulfilled and the number of zero-crossings in each segment can be different, thus the mean and variance can vary.

7.4 Hurst Exponent Estimation

Using different segmentations of the signal, it is possible to determine the crucial statistical characteristics that are dependent on the number of divisions L. In this section we would like to present the key idea of the whole procedure by means of showing the optimal number of non-overlapping subintervals L^*. To formalize such an idea, it is necessary to introduce two requirements:

- unimodality of the $f_{\mathrm{L}}(p)$ function,
- minimization of the number of segments L, where $L > 1$.

The second requirement can be replaced by a stricter condition, that the variance of Hurst exponent estimation has to be minimized. However, considering real data, these two conditions often coincide.

For certain L^* large enough the aggregate probability density function becomes unimodal and we denote this number as the optimal segmentation of the given sample defined as

$$L^* = \begin{cases} 1 & \text{when} f_L(p) \text{ is unimodal} \\ & \text{for all allowable L} \\ \min\{L > 1 : f_L(p) \text{ is unimodal}\} & \text{otherwise} \end{cases}$$

On the basis of the function $f_L^*(p)$ we can determine the boundaries $p_{\min}, p_{\max}$ of confidence interval for p with confidence level $\widetilde{\alpha}$ as

$$\int_0^{p_{\min}} f_{L^*}(p)dp = \frac{\widetilde{\alpha}}{2}, \tag{7.19}$$

$$\int_0^{p_{\max}} f_{L^*}(p)dp = 1 - \frac{\widetilde{\alpha}}{2}. \tag{7.20}$$

Finally, we obtain the boundaries of confidence interval for H as

$$H_{\min} = 1 + \log_2 \cos \frac{p_{\max}\pi}{2}, \tag{7.21}$$

$$H_{\max} = 1 + \log_2 \cos \frac{p_{\min}\pi}{2}, \tag{7.22}$$

where $p_{\min}$ and $p_{\max}$ were calculated from (7.19) and (7.20) with given confidence level $\widetilde{\alpha}$.

In the traditional frequentist method one could calculate the confidence intervals by means of utilizing the properties of standard binomial distribution. Under the assumption that the number of zero-crossings $Z \sim \mathrm{Bi}(N, p^*)$ denote that

$$\mathrm{P}(Z = k) = p_k = \binom{N}{k} (p^*)^k (1 - p^*)^{N-k}. \tag{7.23}$$

One determines the highest value of non-negative integer A and the lowest value of non-negative integer B satisfying conditions

$$\sum_{k=0}^{A} p_k \leqslant \alpha/2, \tag{7.24}$$

$$\sum_{k=B}^{N} p_k \leqslant \alpha/2, \tag{7.25}$$

where α is the confidence level. The adequate boundaries of confidence interval for Hurst exponent equal

$$H_{\min}^{\mathrm{ZC}} = 1 + \log_2 \cos \frac{B\pi}{2N}, \tag{7.26}$$

and

$$H_{\max}^{\mathrm{ZC}} = 1 + \log_2 \cos \frac{A\pi}{2N}, \tag{7.27}$$

where $H_{\min}^{\mathrm{ZC}}$ and $H_{\max}^{\mathrm{ZC}}$ denote the lower and upper bound of the confidence interval, respectively.

Chapter 8

Fractional Analysis of Short Time Series

Previously, we introduced a few methods for analysing fractal sets and signals when we have enough data. It is well known the quality of the estimation significantly depends on the number of input data and with more detailed sampling, the standard deviation of the estimate decreases. There are very few techniques that would be able to extract fractional character from short samples. This is due to several reasons; The first reason is the theory of fractal sets is based on self-similarity and the infinite level of detail that every fractal set has. Most frequently, a large sample is required to be able to reliably capture the density of the points in the space. Second reason is that the methods usually require multiple independent random sampling, which is very often impossible to achieve with real data. Therefore, we aim to present a method that would be able to extract Hurst exponent from short time series based on the autocorrelation properties of the signal.

8.1 Differenced Fractional Brownian Bridge

Inspired by the Brownian Bridge, it is possible to construct fractional generalization of this process that we call fractional Brownian bridge (fBB). The new process employs fBm sample in its definition and can be used in more general models. The construction of this kind of bridge is the key idea for the presented methodology for several reasons:

- the suggested process has still fractional character,
- the process is defined only on fixed interval of finite length, and therefore it is more suitable for short time analysis,
- the fractional bridge starts and ends with zero and therefore, no side effects occur in the case of the left and right zero padding.

Finite sample $B_H(k)$ of size $N+1$ for $k = 0, \ldots, N$ of standardized fBm can be used for the construction of fractional Brownian bridge $M_H(k)$ in the following way

$$M_H(k) = B_H(k) - B_H(0) - \frac{k}{N}(B_H(N) - B_H(0)), \tag{8.1}$$

which is direct generalization of Brownian Bridge. The fBB holds all the important properties of BB such as left and right zero padding [66] and in addition it has desired fractional character. In fractal analysis of time series, the fractional processes are often converted to fractional noises [2] by means of differencing to simplify their covariance structure together with its spectral properties [67] keeping the desired dependence on Hurst exponent. Here we follow this procedure and the final suggested process $X_H(k)$ is presented as the differentiation of fBB. We define the differenced fractional Brownian bridge (dfBB) $X_H(k)$ as

$$X_H(k) = M_H(k+1) - M_H(k) \tag{8.2}$$

for $k = 0, \ldots, N-1$. The sample of fBm with $N = 20$ and corresponding fBB and dfBB are displayed in Fig. 8.1.

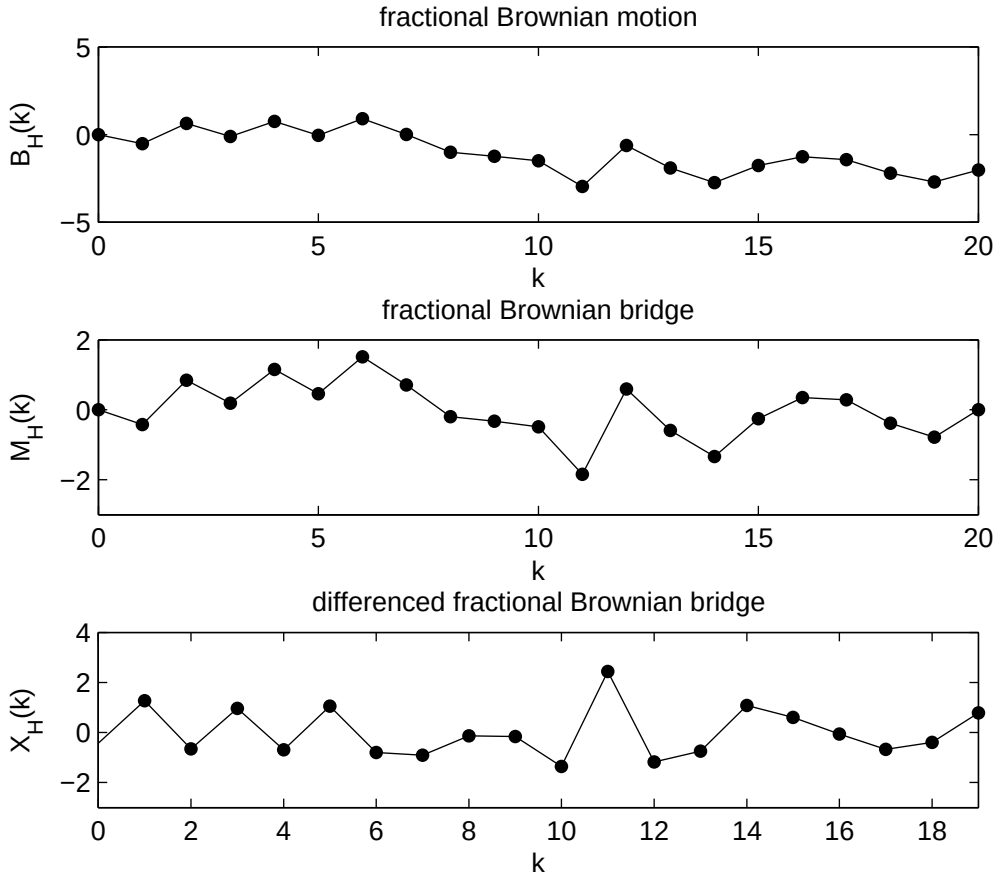


Figure 8.1: Fractional Brownian motion sample and derived processes.

Based on the definition of dfBB we can deduce statistical characteristics that are the subject of following analysis. The process $X_H(k)$ representing differenced fractional Brownian bridge can be simplified as

$$X_H(k) = G_H(k) - \frac{1}{N} B_H(N), \tag{8.3}$$

for $k = 0, 1, \ldots, N-1$ where $G_H(k)$ is sample of fGn and $B_H(N)$ is the last element in the corresponding fBm sequence. The sample $X_H(k)$ itself is fully determined by two parameters – the Hurst exponent H and the length of the sequence N. The appropriate formulas for statistical features are in most cases naturally dependent on these two parameters.

Due to the fact, that the dfBB comprises of two processes with both zero expected value, the expected value of dfBB equals zero too, i.e

$$\mathrm{E}(X_H(k)) = \mathrm{E}(G_H(k)) - \frac{1}{N}\mathrm{E}(B_H(N)) = 0. \tag{8.4}$$

A little bit more difficult can be the computation of the variance $\mathrm{E}(X_H^2(k))$. Using the autocovariance structure of fBm given in (2.13) and rewriting fGn as difference of fBm from definition (2.14)one obtains

$$\begin{aligned} \mathrm{E}(X_H^2(k)) =& \mathrm{E}(B_H^2(k+1) + B_H^2(k) + \frac{1}{N^2}B_H^2(N) - \\ & 2B_H(k+1)B_H(k) - \\ & \frac{2}{N}B_H(k+1)B_H(N) + \\ & \frac{2}{N}B_H(k)B_H(N)). \end{aligned} \tag{8.5}$$

Using property (2.13) of $B_H(k)$ we obtain

$$\begin{aligned} \mathrm{E}(X_H^2(k)) =& 1 - N^{2H-2} + \\ & + \frac{2}{N}(|k|^{2H} - |k+1|^{2H} + |k-N|^{2H} - |k+1-N|^{2H}). \end{aligned} \tag{8.6}$$

Finally, the variance of the dfBB process can be subsequently written as

$$\gamma_0 = \mathrm{var}(X_H) = \frac{1}{N}\sum_{k=0}^{N-1} \mathrm{E}(X_H^2(k)) = 1 - N^{2H-2}. \tag{8.7}$$

Calculation of autocovariance requires the computation of $\mathrm{E}(X_H(k)X_H(l))$ for $k, l = 0, \ldots, N-1$ where $k > l$. The expected value of $X_H(k)X_H(l)$ can be expressed as

$$\begin{aligned} & \mathrm{E}(X_H(k)X_H(l)) = \\ & \mathrm{E}\left(G_H(k)G_H(l) - \frac{1}{N}G_H(l)B_H(N)\right) - \\ & \mathrm{E}\left(\frac{1}{N}G_H(k)B_H(N) + \frac{1}{N^2}B_H^2(N)\right). \end{aligned} \tag{8.8}$$

Using the correlation between G_H and B_H as

$$\mathrm{E}(G_H(k)B_H(l)) = \frac{1}{2}\left(|k+1|^{2H} - |k|^{2H} + |k-l|^{2H} - |k+1-l|^{2H}\right) \tag{8.9}$$

together with substitution $m = k - l$, one obtains that the autocovariance is

$$\gamma_m = \eta(m, H) + N^{2H-2} + \frac{|m|^{2H} - |N-m|^{2H} - |N|^{2H}}{N(N-m)}, \tag{8.10}$$

that is expressed as a sum of fGn autocorrelation function η and correction of order $\mathcal{O}\left(N^{2H-2}\right)$ as N goes to infinity. The corresponding autocorrelation function is expressed as

$$\rho_m = \frac{\gamma_m}{\gamma_0} \tag{8.11}$$

for $m = 0, \ldots, N-1$.

8.2 Parameter Estimation Method

Having generated $Q \in \mathbb{N}$ independent samples of fGn of length $N \in \mathbb{N}$ with the same theoretical value of Hurst exponent H, the new methodology of estimation could be applied to investigate fractional properties of these trajectories. By means of cumulative sum, the fGn sample is transformed into fBm and utilizing formula (8.1), the fBB can be constructed. The dfBB sample $x_0, x_1, \ldots, x_{M-1}$ arises as a differentiation of fBB. The estimation of n-th autocovariance coefficient $\widehat{r}_n$ can be expressed for $n = 0, \ldots, N-1$ as

$$\widehat{r}_n = \sum_{k=0}^{M-n-1} x_k x_{k+n} \tag{8.12}$$

and the estimation of autocorrelation coefficient $\widehat{\rho}_n$ as

$$\widehat{\rho}_n = \frac{\widehat{r}_n}{\widehat{r}_0}. \tag{8.13}$$

Denote the theoretical value of autocorrelation from equation (8.11) as $\rho_n = \rho_n(H)$ and the averaged autocorrelation from Q independent realisations as $\mathrm{E}(\widehat{\rho}_n)$. Than we obtain the estimation of parameter H by means of solving the minimization problem

$$\widehat{H} = \underset{H\in(0,1)}{\operatorname{argmin}} \sum_{n=1}^{N-1} \left(\mathrm{E}(\widehat{\rho}_n) - \rho_n(H)\right)^2. \tag{8.14}$$

The point estimate of $\widehat{H}$ can be obtained by the maximum likelihood method [68] together with its standard deviation.

Chapter 9

Hurst Exponent Estimation via Wishart Distribution

Autocorrelation properties of fractional processes are very useful for estimating their Hurst exponent. The idea of the following approach is following: Having independent samples of the same process allows to create a structure that would capture and generalize the autocorrelation properties of the original process. We assume that we have a couple of samples of n-dimensional process and will form a scatter matrix, that summarizes the properties of all samples. Using the Wishart distribution and the summarized matrix, it is possible to derive the statistical properties of the original process. We will demonstrate both on how to apply this method on one-dimensional signals as well as on multidimensional structures, such as fractional surfaces and fractional masses.

9.1 Wishart Distribution Primer

This section summarizes the key properties of Wishart distribution [69] that will be applied later. The Wishart distribution generalizes gamma distribution and it is defined over random matrices. Since in this work, we consider mostly Gaussian processes [70], we use finite length samples of fractional processes as samples from a multidimensional normal distribution with given covariance matrix $\mathbb{V}$.

Let $p \in \mathbb{N}, n \in \mathbb{N}$ be number of dimensions and samples, respectively. Let $\mathbb{V} \in \mathbb{R}^{p\times p}, \mathbb{V} > 0$ be covariance matrix. Let $\mathbb{G} = (\vec{g}_1, \vec{g}_2, \dots \vec{g}_n) \in \mathbb{R}^{p\times n}$ be the sample matrix, where $\vec{g_j} \sim \mathrm{N}(\vec{0}, \mathbb{V})$. Then the scatter matrix $\mathbb{S} = \mathbb{G}\mathbb{G}^{\mathrm{T}} \in \mathbb{R}^{p\times p}$ has Wishart distribution, i.e. $\mathbb{S} \sim \mathrm{W}(\mathbb{V}, p, n)$.

Adequate probability density function (pdf) can be expressed as

$$f(\mathbb{S}, \mathbb{V}, p, n) = \frac{|\mathbb{S}|^{(n-p-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbb{V}^{-1}\mathbb{S})\right)}{2^{np/2}\,|\mathbb{V}|^{n/2}\,\Gamma_p(\frac{n}{2})}, \tag{9.1}$$

where

$$\Gamma_p(\xi) = \pi^{p(p-1)/2} \cdot \prod_{j=1}^{p} \Gamma\left(\xi - \frac{j-1}{2}\right) \tag{9.2}$$

and tr(...) means trace of a matrix.

There is a useful property of Wishart distribution. When the transform matrix $\mathbb{C} \in \mathbb{R}^{q \times p}$ has rank equal to q, then

$$\mathbb{C}\mathbb{S}\mathbb{C}^{\mathrm{T}} \sim \mathrm{W}(\mathbb{C}\mathbb{V}\mathbb{C}^{\mathrm{T}}, q, n). \tag{9.3}$$

The previous property is useful for the analysis of transformed processes. Given a sample matrix $\mathbb{G}$, we can transform the data using a transformation matrix $\mathbb{C}$, thus the new process sample matrix will equal $\mathbb{G}_{\mathrm{new}} = \mathbb{C}\mathbb{G}$ and resulting scatter matrix equals $\mathbb{G}_{\mathrm{new}}\mathbb{G}_{\mathrm{new}}^{\mathrm{T}} = \mathbb{C}\mathbb{S}\mathbb{C}^{\mathrm{T}}$. Therefore, for any derived process for which we know the transformation matrix, we immediately know the parameters of Wishart distribution as well.

When the transformation matrix $\mathbb{C}$ has full rank, we obtain

$$\mathrm{tr}(\mathbb{V}_{\mathrm{new}}^{-1} \cdot \mathbb{S}_{\mathrm{new}}) = \mathrm{tr}(\mathbb{V}^{-1}\mathbb{S}), \tag{9.4}$$

where $\mathbb{S}_{\mathrm{new}} = \mathbb{C}\mathbb{S}\mathbb{C}^{\mathrm{T}}$ and $\mathbb{V}_{\mathrm{new}} = \mathbb{C}\mathbb{V}\mathbb{C}^{\mathrm{T}}$. Therefore, the pdf of transformed process is only rescaled pdf of the original one. In the particular case, when $\mathbb{C}$ is square matrix with $|\mathbb{C}| = 1$, the pdf remains unchanged.

We will relate the previously introduced terms to the fractional random processes. Fractional Brownian motion of length $N \in \mathbb{N}$ is a Gaussian process dependent on parameters $H \in (0; 1), \sigma > 0$, therefore

$$\vec{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}} \sim \mathrm{N}\left(\vec{0}, \mathbb{V}_0\right), \tag{9.5}$$

where $\mathbb{V}_0 = (v_{i,j})_{i,j=0}^{N}$ for

$$v_{i,j} = \frac{\sigma^2}{2}\left(|i|^{2H} + |j|^{2H} - |i-j|^{2H}\right). \tag{9.6}$$

There are various stochastic processes, that are derived from fBm. Frequently used fractional Gaussian noise (fGn) generates a sample $(y_i)_{i=1}^{N}$ as fBm differentiation as

$$y_i = x_i - x_{i-1}, \tag{9.7}$$

which can be rewritten as

$$\vec{y} = \mathbb{C}\vec{x} \tag{9.8}$$

where $\mathbb{C} = (c_{i,j})_{i,j=1}^{N}$ and $c_{i,j} = \delta_{i,j} - \delta_{i,j-1}$ for Kronecker delta $\delta_{i,j}$. Resulting matrix is triangular and $|\mathbb{C}| = 1$. Adequate covariance matrix of fGn is $\mathbb{V} = \mathbb{C}\mathbb{V}_0$.

Another possibility is to analyse fractional bridge. The fractional Brownian bridge (fBB) generates a sample $(z_i)_{i=0}^{N}$ defined as

$$z_i = x_i - x_0 - \frac{i}{N}\left(x_N - x_0\right). \tag{9.9}$$

The transformation matrix $\mathbb{C} \in \mathbb{R}^{(N-1)\times N}$ has decreased rank and adequate covariance matrix is $\mathbb{V} = \mathbb{C}\mathbb{V}_0$.

During the fractal analysis of time series, the fractional processes are often converted to fractional noises by means of differencing to simplify their covariance structure together with its spectral properties, while keeping the desired dependence on Hurst exponent. The sample $(w_i)_{i=1,N}$ of differenced fractional Brownian bridge (dfBB) is defined as

$$w_i = z_i - z_{i-1} \tag{9.10}$$

and $\mathbb{C} \in \mathbb{R}^{(N-1)\times N}, \mathbb{V} = \mathbb{C}\mathbb{V}_0$ again.

All mentioned discrete processes and derived samples are dependent on the Hurst exponent H determining whether the process has long or short memory.

9.2 Maximum Likelihood Estimation Approach

Observing independent fBm samples of length N, we can compare the experimental scatter matrix with theoretical covariance matrix using maximum likelihood approach [71]. The scatter matrix can be evaluated from $n > N$ samples of fBm as

$$\mathbb{S}_0 = \sum_{k=1}^{n} \vec{x}_k \vec{x}_k^{\mathrm{T}}, \tag{9.11}$$

which is single, but summarized result of the fBm observations. The likelihood value is therefore

$$L(H, \sigma) = f(\mathbb{S}_0, \mathbb{V}_0, N, n), \tag{9.12}$$

because $\mathbb{V}_0$ is function of H, σ. Using adequate negative-logarithm transformation, the likelihood maximization can be converted to a minimization of

$$\Phi(H, \sigma) = -\ln L(H, \sigma) = Q + \frac{n}{2} \ln |\mathbb{V}_0| + \frac{1}{2} \mathrm{tr}\left(\mathbb{V}_0^{-1} \mathbb{S}_0\right), \tag{9.13}$$

where Q is dependent on the input data, but independent on H, σ. The range of H is constrained, but $\sigma > 0$ could vary in order. Therefore, we prefer to use $r = \ln \sigma$ as more suitable parameter. Resulting optimization task is

$$\Psi(H, r) = \Phi(H, \exp(r)) \to \min. \tag{9.14}$$

The optimum values H^*, r^* are the point estimates of H, r in the sense of likelihood technique. After the evaluation of Hessian matrix [71]

$$\mathbb{H} = \left(\begin{array}{cc} \frac{\partial^2 \Psi}{\partial H^2} & \frac{\partial^2 \Psi}{\partial H \partial r} \\ \frac{\partial^2 \Psi}{\partial r \partial H} & \frac{\partial^2 \Psi}{\partial r^2} \end{array} \right)_{H=H^*, r=r^*} > 0 \tag{9.15}$$

we directly evaluate asymptotic variance of H^* estimate as $u_{1,1}$ where $\mathbb{U} = \mathbb{H}^{-1}$ and adequate standard deviation is obtained as $sd = \sqrt{u_{1,1}}$. The unbiasedness of

proposed method can be measured by z-score as $z = (H - H^*)/sd$, which should satisfy $|z| \leqslant 1.960$ for significance level 0.05. The corresponding 95% confidence interval is therefore

$$\mathrm{CI} = [H^* - 1.96sd; H^* + 1.96sd]. \tag{9.16}$$

The confidence region for parameter pair (H, r) is defined as set

$$\mathrm{CR} = \left\{ (H, r) : \Psi(H, r) \leqslant \Psi(H^*, r^*) + \frac{\chi^2_2(0.95)}{2} \right\}. \tag{9.17}$$

We can calculate the mean square error as precision measure

$$MSE = sd \cdot sd + (H^* - H)^2. \tag{9.18}$$

Similar methodology can be used also for derived processes. Using the transform matrix $\mathbb{C}$, we can obtain another covariance matrix $\mathbb{V} = \mathbb{C}\mathbb{V}_0\mathbb{C}^{\mathrm{T}}$ of a given process. The Corresponding scatter matrix can be written as

$$\mathbb{S} = \sum_{k=1}^{n} \vec{y}_k \vec{y}_k^{\mathrm{T}} = \mathbb{C}\mathrm{S}_0\mathbb{C}^{\mathrm{T}}. \tag{9.19}$$

After the application of maximum likelihood technique, we have

$$\Phi(H, \sigma) = -\ln L(H, \omega) = Q + \frac{n}{2} \ln |\mathbb{V}| + \frac{1}{2} \mathrm{tr}\left(\mathbb{V}^{-1}\mathbb{S}\right). \tag{9.20}$$

Finding minima of $\Phi(H, \sigma)$ at point (H^*, r^*), we can calculate all other statistical properties, such as sd, CI, CR and MSE.

9.3 Generalized Circulant Embedding

The methods for simulation of one-dimensional fBm are known, but the generation methods of multidimensional fBm from (2.24) are not well developed. We will need an exact method for multidimensional fBm generation to be able to verify its properties via simulations. This section describes an idea of multidimensional fBm generation that is based on the generalization of traditional circulant embedding [72, 73] (TCE) method. We introduce the general circulant embedding (GCE) method as follows. The method utilizes the simulation of incremental fractional Gaussian noise (ifGn) signal in its core. Suppose that M is even size of n-dimensional integer grid $\mathcal{D} = \{-M/2, \ldots, M/2 - 1\}^n$ which is a support of $B_H(\vec{k})$ as fBm and $Y_H(\vec{k})$ as ifGn for $\vec{k} \in \mathcal{D}$.

Using index vector $\vec{k} \in \mathcal{D}$, we will study the relationship between $B_H(\vec{k})$ and $Y_H(\vec{k})$. Using discrete Laplacian Δ_D in cubic domain, we can directly calculate

$$Y_H(\vec{k}) = \Delta_D B_H(\vec{k}) = \sum_{\substack{\vec{u} \in \mathbb{Z}^n \\ \|\vec{u}\| = 1}} \left(B_H(\vec{k} + \vec{u}) - B_H(\vec{k}) \right). \tag{9.21}$$

Having $B_H(\vec{0}) = 0$, we can continue with

$$C(\vec{k}) = \mathrm{E}\left(Y_H((\vec{k})Y_H(\vec{0}\right) \tag{9.22}$$

$$= \mathrm{E}\left(\sum_{\substack{\vec{u}\in\mathbb{Z}^n \\ \|\vec{u}\|=1}} \left(B_H(\vec{k}+\vec{u}) - B_H(\vec{k})\right) \cdot \sum_{\substack{\vec{v}\in\mathbb{Z}^n \\ \|\vec{v}\|=1}} \left(B_H(\vec{v}) - B_H(\vec{0})\right)\right) \tag{9.23}$$

$$= \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \mathrm{E}\left(B_H(\vec{k}+\vec{u})B_H(\vec{v}) - B_H(\vec{k})B_H(\vec{v})\right) \tag{9.24}$$

$$= \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \mathrm{E}\left(B_H(\vec{k}+\vec{u})B_H(-\vec{v}) - B_H(\vec{k})B_H(-\vec{v})\right) \tag{9.25}$$

$$= \frac{\sigma^2}{2} \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \Big(\|\vec{k}+\vec{u}\|^{2H} + \|-\vec{v}\|^{2H} - \|\vec{k}+\vec{u}+\vec{v}\|^{2H} - \tag{9.26}$$

$$- \|\vec{k}\|^{2H} - \|-\vec{v}\|^{2H} + \|\vec{k}+\vec{v}\|^{2H}\Big) =$$

$$= \frac{\sigma^2}{2} \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \left(\|\vec{k}\|^{2H} - \|\vec{k}+\vec{u}\|^{2H} - \|\vec{k}+\vec{v}\|^{2H} + \|\vec{k}+\vec{u}+\vec{v}\|^{2H}\right). \tag{9.27}$$

Using matrix $\mathbb{W}_H \in \mathbb{R}^{M^n}$ of elements $W_H(\vec{k}) = \|\vec{k}\|^{2H}$ for $\vec{k} \in \mathcal{D}$ we can also apply the discrete Laplacian Δ_D to produce

$$\Delta_D W_H(\vec{k}) = \sum_{\substack{\vec{u}\in\mathbb{Z}^n \\ \|\vec{u}\|=1}} \left(W_H(\vec{k}+\vec{u}) - W_H(\vec{k})\right) \tag{9.28}$$

and also the convolution

$$\Big(\Delta_D * \Delta_D\Big) W_H(\vec{k}) = \tag{9.29}$$

$$= \sum_{\substack{\vec{v}\in\mathbb{Z}^n \\ \|\vec{v}\|=1}} \left(\Delta_D W_H(\vec{k}+\vec{v}) - \Delta_D W_H(\vec{k})\right) = \tag{9.30}$$

$$= \sum_{\substack{\vec{v},\vec{u}\in\mathbb{Z}^n \\ \|\vec{v}\|=\|\vec{u}\|=1}} \left(W_H(\vec{k}+\vec{u}+\vec{v}) - W_H(\vec{k}+\vec{u}) - W_H(\vec{k}+\vec{v}) + W_H(\vec{k})\right) = \tag{9.31}$$

$$= \sum_{\substack{\vec{v},\vec{u}\in\mathbb{Z}^n \\ \|\vec{v}\|=\|\vec{u}\|=1}} \left(\|\vec{k}+\vec{u}+\vec{v}\|^{2H} - \|\vec{k}+\vec{u}\|^{2H} - \|\vec{k}+\vec{v}\|^{2H} + \|\vec{k}\|^{2H}\right). \tag{9.32}$$

Therefore,

$$C(\vec{k}) = -\frac{\sigma^2}{2} \left(\Delta_D * \Delta_D\right) W_H(\vec{k}) \tag{9.33}$$

or in matrix form as

$$\mathbb{C} = -\frac{\sigma^2}{2}\Big(\Delta_D * \Delta_D\Big)\mathbb{W}_H. \tag{9.34}$$

Process Y_H is still Gaussian with known autocovariance structure. Moreover, its autocorrelation is not dependent on the time lag, therefore it can be generated using the circulant embedding method.

Having randomly generated sample of ifGn as $Y_H(\vec{k})$ with zero mean value, we recognize (9.21) as Discrete Poisson Equation [74] for an instance of fBm $B_H(\vec{k})$ on the integer domain $\mathcal{D}$. Denoting $\mathcal{DFT}$, $\mathcal{DFT}^{-1}$ as Discrete Fourier transform and its inversion on $\mathcal{D}$, we can directly solve (9.21) with periodic boundary conditions. In this case, (9.21) has infinite many solutions which differ by constants.Without loosing generality, we will find the solution with zero mean value as follows. First, we will apply the discrete Fourier transform to (9.21) and obtain

$$\mathcal{DFT}\{Y_H(\vec{k})\} = F(\vec{\omega}) \otimes \mathcal{DFT}\{B_H(\vec{k})\},$$

where

$$F(\vec{\omega}) = -4 \sum_{j=1}^{n} \sin^2(\omega_j/2) \in \mathbb{R}$$

with $\vec{\omega} = 2\pi\vec{k}/N$ for $\vec{k} \in \mathcal{D}$ is transformed discrete Laplacian operator and $\otimes$ is component–wise multiplication. The deconvolution can be obtained using Wiener filter [75, 76] with parameter $\lambda > 0$ using formula

$$B_H(\vec{k}) = \mathcal{DFT}^{-1}\{W(\vec{\omega}) \otimes \mathcal{DFT}\{Y_H(\vec{k})\}\}, \tag{9.35}$$

where

$$W(\vec{\omega}) = \frac{F^*(\vec{\omega})}{\lambda^2 + |F(\vec{\omega})|^2} = \frac{F(\vec{\omega})}{\lambda^2 + F^2(\vec{\omega})}. \tag{9.36}$$

9.4 Multi-Dimensional Generalization

Recall n-dimensional fBm denoted as $B_H(\vec{x})$ satisfies

$$\mathrm{E}\left(B_H(\vec{x})B_H(\vec{y})\right) = \frac{\sigma_0^2}{2} \cdot \left(\|\vec{x}\|^{2H} + \|\vec{y}\|^{2H} - \|\vec{x} - \vec{y}\|^{2H}\right) \tag{9.37}$$

and $\mathrm{E}\, B_H(\vec{x}) = 0$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ where $\|..\|_2$ denotes Euclidean distance. Using rectangular equidistant sampling with step $h > 0$ and denoting $B(\vec{K}) = B_H(h\vec{K})$ for $\vec{K} \in \mathbb{Z}^n$, we obtain discrete fBm formula satisfying $\mathrm{E}\, B(\vec{K}) = 0$ and

$$\mathrm{E}\left(B(\vec{K})B(\vec{L})\right) = \frac{\sigma^2}{2}\left(\|\vec{K}\|^{2H} + \|\vec{L}\|^{2H} - \|\vec{K} - \vec{L}\|^{2H}\right) \tag{9.38}$$

for $\vec{K}, \vec{L} \in \mathbb{Z}^n$, where $\sigma = \sigma_0 \cdot h^H$.

The discretized n-dimensional signal can be sampled by symmetric compact mask

$$\mathcal{M}^* = \{\vec{s} \in \mathbb{Z}^n : \|\vec{s}\|_p \leqslant \rho\} \tag{9.39}$$

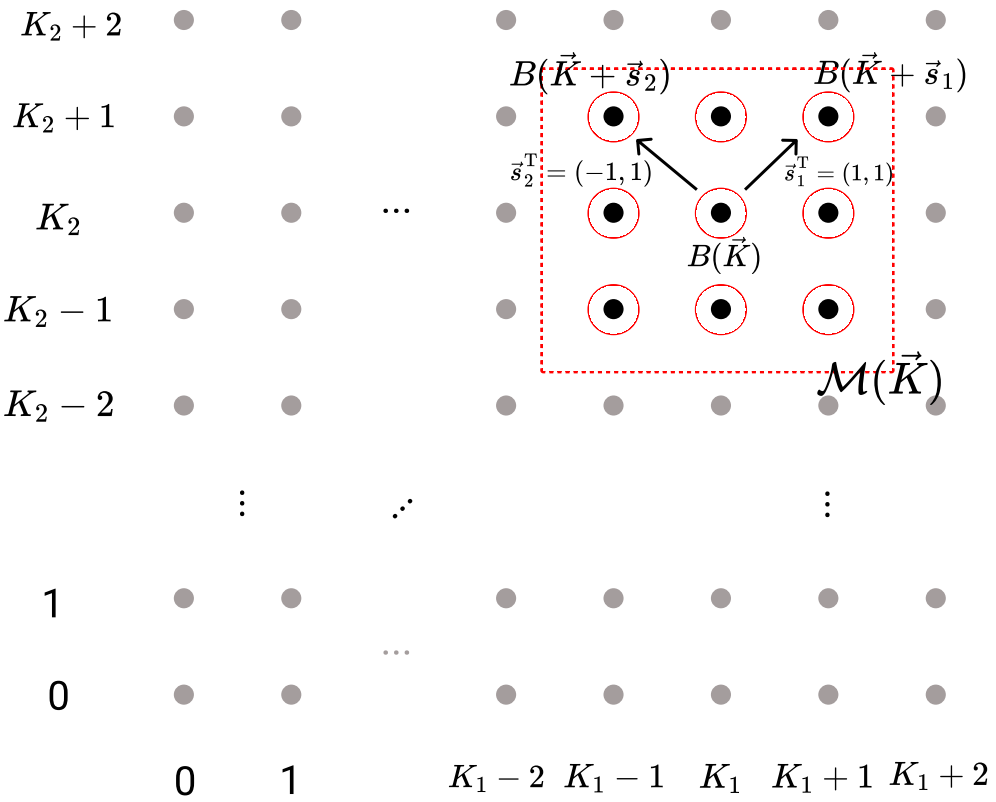


Figure 9.1: Example of rectangular mask applied to two-dimensional fBm.

using Minkowski norm $\|..\|_p$ for $p, \rho \geqslant 1$. The parameter ρ is later called mask radius. The mask $\mathcal{M}^*$ consists of $N+1$ points $\vec{s}_0 = \vec{0}, \vec{s}_1, \ldots, \vec{s}_N$ and can be shifted from the origin to a point $\vec{K} \in \mathbb{Z}^n$ to form

$$\mathcal{M}(\vec{K}) = \left\{ \vec{K} + \vec{s} : \vec{s} \in \mathcal{M}^* \right\}. \tag{9.40}$$

Therefore, the sample at point $\vec{K}$ consists of values $B(\vec{K}), B(\vec{K}+\vec{s}_1), \ldots, B(\vec{K}+\vec{s}_N)$. The neighbourhood of $B(\vec{K})$ and its generation for rectangular mask with nine elements is visualized in Fig. 9.1. We investigate the reduced sample of size N and values $\xi_i = B(\vec{K}+\vec{\xi}_i) - B(\vec{K})$ for $i = 1, \ldots, N$. Adequate covariance matrix $\mathbb{V} = \operatorname{var} \vec{x}_i$ has elements

$$v_{i,j} = \mathrm{E}(\xi_i \xi_j) = \mathrm{E}\left(B(\vec{K}+\vec{s}_i) - B(\vec{K})\right)\left(B(\vec{K}+\vec{s}_j) - B(\vec{K})\right) = \tag{9.41}$$

$$= \mathrm{E}\left(B(\vec{s}_i) - B(\vec{0})\right)\left(B(\vec{s}_j) - B(\vec{0})\right) = \tag{9.42}$$

$$= \frac{\sigma^2}{2} \cdot \left(\|\vec{s}_i\|^{2H} + \|\vec{s}_j\|^{2H} - \|\vec{s}_i - \vec{s}_j\|^{2H}\right) \tag{9.43}$$

which is independent on $\vec{K}$. The theoretical covariance matrix of the sample has the same dependences as the covariance matrix of the original process.

Given n-dimensional fBm and mask $\mathcal{M}^*$, we consider m sets $\mathcal{M}(\vec{r}_q)$ where $\vec{r}_q \in \mathbb{Z}^n$ for $q = 1, \ldots, m$. The parameter m will be further called number of samples. For

each set, we define sample vector $\vec{g}_q \in \mathbb{R}^N$ as

$$\vec{g}_q^{\mathrm{T}} = (B(\vec{r}_q + \vec{s}_1) - B(\vec{r}_q), B(\vec{r}_q + \vec{s}_2) - B(\vec{r}_q), \ldots, B(\vec{r}_q + \vec{s}_N) - B(\vec{r}_q)) \,. \quad (9.44)$$

for $q = 1, \ldots, m$. Let $\mathbb{G} = (\vec{g}_1, \vec{g}_2, \ldots, \vec{g}_m) \in \mathbb{R}^{N \times m}$ be the sample matrix, having $\vec{g}_q \sim \mathrm{N}(\vec{0}, \mathbb{V})$. The scatter matrix $\mathbb{S} = \mathbb{G}\mathbb{G}^{\mathrm{T}}$ has Wishart distribution [69] with probability density function

$$f(\mathbb{S}, \mathbb{V}, N, m) = \frac{|\mathbb{S}|^{\frac{m-N-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbb{V}^{-1}\mathbb{S})\right)}{2^{mN/2} \, |\mathbb{V}|^{m/2} \, \Gamma_N(\frac{m}{2})}, \quad (9.45)$$

where

$$\Gamma_N(\xi) = \pi^{N(N-1)/2} \cdot \prod_{j=1}^{N} \Gamma\left(\xi - \frac{j-1}{2}\right) \quad (9.46)$$

and tr(...) means trace of a matrix.

Given $m > N$ samples of n-dimensional fBm, one can calculate the scatter matrix as

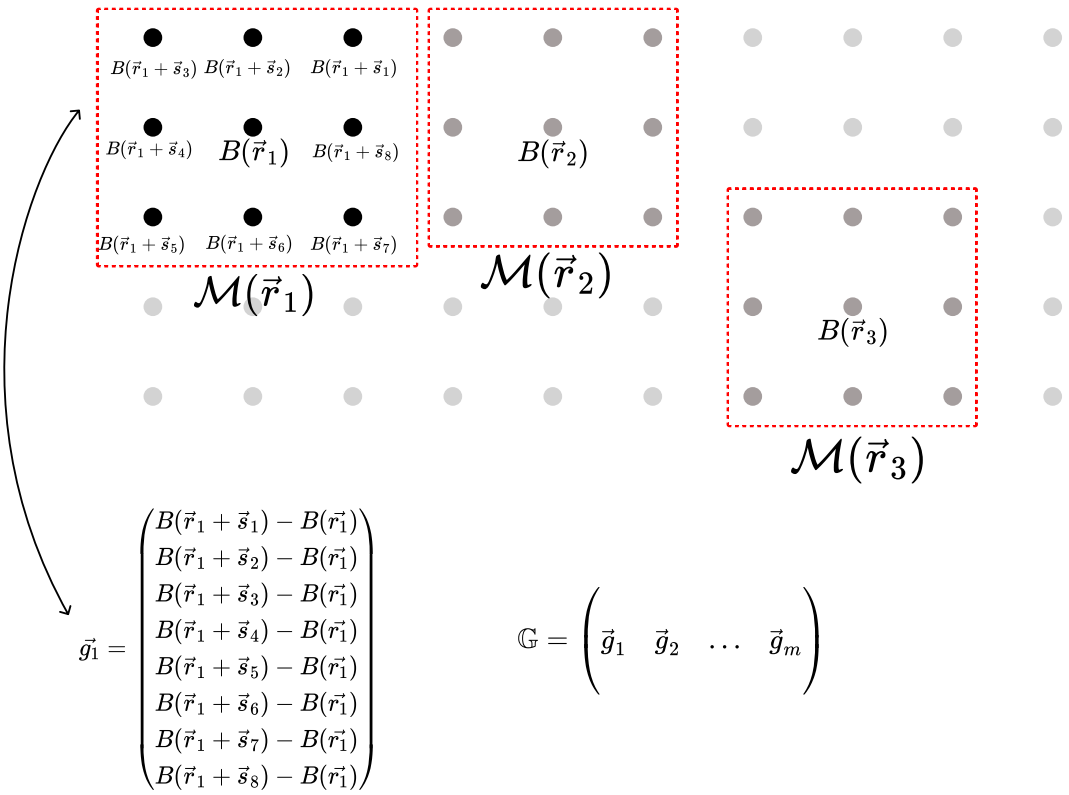
$$\mathbb{S} = \sum_{l=1}^{m} \vec{g}_j \vec{g}_j^{\mathrm{T}}. \quad (9.47)$$

The scatter matrix generation from two-dimensional fBm and rectangular mask is visualized in Fig. 9.2.

By maximizing the value of corresponding likelihood

$$L(H, \sigma) = f(\mathbb{S}, \mathbb{V}(H, \sigma), N, m) \to \max, \quad (9.48)$$

one gets the optimum estimated values $(\widehat{H}, \widehat{\sigma})$ along with the standard deviations of the estimates. For standardized fBm with $\sigma = 1$, we will use the mean square error (MSE) as a measure of accuracy.



$$\vec{g}_1 = \begin{pmatrix} B(\vec{r}_1+\vec{s}_1)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_2)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_3)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_4)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_5)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_6)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_7)-B(\vec{r}_1) \\ B(\vec{r}_1+\vec{s}_8)-B(\vec{r}_1) \end{pmatrix} \qquad \mathbb{G} = \begin{pmatrix} \vec{g}_1 & \vec{g}_2 & \cdots & \vec{g}_m \end{pmatrix}$$

Figure 9.2: Scatter matrix generation from two-dimensional fBm sample.

Chapter 10

Statistical Testing on Simulated Data

Having described the theoretical properties of the estimation algorithms, this chapter summarizes the results on how the methods perform when applied to simulated data with known fractal dimension. Here we describe on how we generate the training sets and how we validate the method performance. We will be verifying following methods on simulated data sets:

- Rotational spectrum for D_2 estimation,
- Modified Renyi entropy for D_α estimation,
- Short time series methodology for H estimation,
- Wishart distribution method for H estimation.

The rotational spectrum and Renyi entropy use as input point sets, therefore we summarize the sets with known dimension used within this chapter in Tab. 10.1. The other two methods use fBm as input, and therefore we summarize techniques for generating fBm in Tab. 10.2.

The methods are tested for wide range of input parameters and eligible dimensions as well as for their input parameters. The outcome of the simulation is verification, whether the methods works as expected, how accurate it is and what is the recommended parameter settings to achieve best results. The methods have usually internal parameters that directly influence the estimation routine (e.g. lower and upper bound for performing regression, size of segments) as well as minimum requirement for the number of data points. Each of the method is tested on fractals with known dimension and the optimal setting is recommended. Thorough this chapter, we will use standard notation that is summarized in Tab. 10.3.

Table 10.1: Sets with known dimension.

structure name	parameter(s)	meaning	reference
Cantor set	$a = 1/3$	contraction ratio	[17]
Cantor dust	$a = 1/3$	contraction ratio	[77]
Sierpinski carpet	$a = 1/3$	contraction ratio	[18]
Real numbers with even digits set	-	-	[3]
Takagi function	$a \in (1/2, 1)$	scale ratio	[22]
	$n \in \mathbb{N}$	dimension	
fBm trajectory	$H \in (0, 1)$	Hurst exponent	[33, 34]
	$n \in \mathbb{N}$	space dimension	
Levy flight	$\alpha \in (0, 2)$	stability parameter	[31]
	$n \in \mathbb{N}$	dimension	
Henon map	$a = 0.4, b = 0.3$	system parameters	[78]
	$(x_0 = 0, y_0 = 0.9)$	starting point	
Lozi map	$a = 1.7, b = 0.5$	system parameters	[79]
	$(x_0 = -0.1, y_0 = 0.1)$	starting point	
de Wijs' fractal	$a \in (0, 1/2)$	system parameter	[19]

Table 10.2: Methods for fBm estimation.

method name	range of H	space	reference
Lowen method	$(0, 0.5]$	1D	[39]
Davies-Harte method	$(0, 1)$	1D	[41, 45]
Abry-Sellan method	$(0.5, 1)$	1D	[46]
Traditional circulant embedding	$(0, 1)$	1D	[80]
Generalized circulant embedding	$(0, 1)$	nD	section 9.3

Table 10.3: Notation of variables.

variable	meaning
H,D_α	theoretical value of Hurst exponent (Renyi dimension)
$\widehat{H}$,$\widehat{D_\alpha}$	point estimate of Hurst exponent (Renyi dimension)
sd	standard deviation of the estimate
z	z-score (unbiasedness of the estimate), denoted as ** if lower than 1.24 and * if lower than 1.96
p-value	probability that the point estimate does not differ from the theoretical value

10.1 Estimating D_2 using Rotational Spectrum

The methodology of estimating correlation dimension via rotational spectrum employs its dependence on the logarithm of Ω as

$$\lim_{\Omega\to\infty} \frac{\ln S(\Omega)}{\ln \Omega} = \lim_{\Omega\to\infty} \frac{\log S(\Omega)}{\log \Omega} = -D_2. \tag{10.1}$$

and can be estimated using the linear model

$$\log S(\Omega) = A - D_2 \cdot \log\ \Omega + \epsilon \tag{10.2}$$

for suitable values Ω. The aim is to validate that the estimation of correlation dimension is unbiased for sets with known dimension and recommend parameters $\Omega_{\min}, \Omega_{\max}$ for regression alongside a suitable kernel function. Recall the rotational spectrum of a finite sample is defined as

$$\widehat{S}(\Omega) = \frac{1}{M} \sum_{j=1}^{M} H_n\left(\Omega \|\vec{x}_j - \vec{y}_j\|\right) \tag{10.3}$$

for M point pairs $\vec{x}_j, \vec{y}_j \in \mathcal{F}$ for $j = 1, \ldots, M$ and $n \in \mathbb{N}$.

The main feature of the proposed methodology is its smoother dependence of the spectrum on Ω in comparison to traditional approach with correlation sum (3.14). The Fig. 10.1 compares the dependency of correlation sum $C(r)$ on the distance r and rotational spectrum $S(\Omega)$ with kernel function H_∞ on the frequency Ω. The investigated set is Cantor dust with $M = 10^5$ point pairs. Subjectively one can argue that the rotational spectrum has much more smooth dependency in comparison to the correlation sum, therefore could carry more precise information of the correlation dimension. For its estimation, we will be using L_p regression with $p = 4$ while using the model (10.2).

At first, the dependency of variance of the estimation on the total number of point pairs M is investigated. The reference set is again Cantor dust with various number of generated point pairs. Table 10.4 presents the point estimate its variance depending on the number of points M. Rotational spectrum is calculated for kernel H_∞. Both in the case of correlation sum and in the case of rotational spectrum, the standard deviation *sd* of the estimate decreases, though for $M > 10^4$, the rotational spectrum estimation provides the same accurate estimate, but with smaller standard deviation.

It is also possible to estimate the rotational spectrum for finite dimensional rotation using the kernel functions H_n for $n \in N$. We have proven that the correlation dimension estimation coincides with D_2 in case of infinite rotational spectrum, but we will validate experimentally that the relationship holds also for other degrees of rotation $n \in \mathbb{N}$. The comparison of the kernel functions that can be used for the rotation of the power spectrum is shown in Fig. 10.2 for H_2, H_3, H_4 and H_∞. The traditional Sierpinski carpet was used for this simulation.

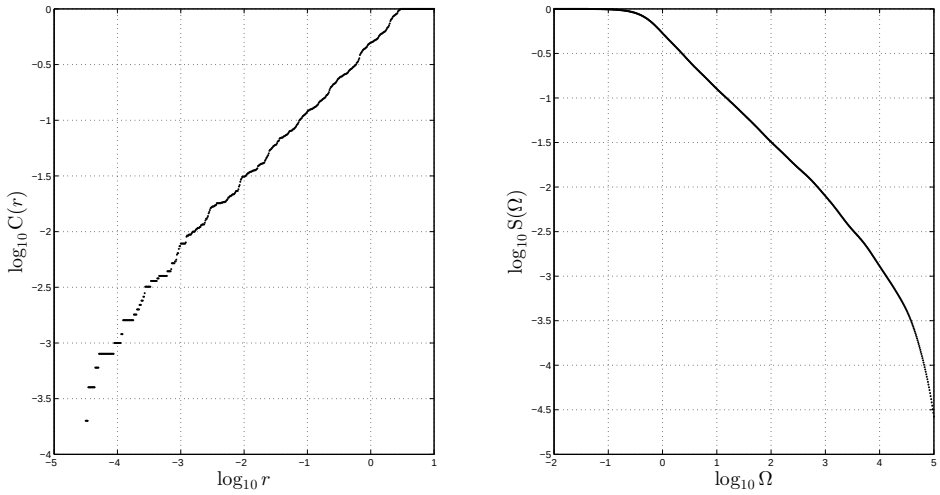


Figure 10.1: Traditional correlation sum (left) and rotational spectrum (right).

Table 10.4: Cantor dust analysis using L_4 regression.

	Correlation sum				Rotational spectrum			
M	$\widehat{D_2}$	sd	z	p-value	$\widehat{D_2}$	sd	z	p-value
10^3	1.2941	0.1178	**	0.7843	1.2378	0.1010	**	0.8117
10^4	1.2937	0.0803	**	0.6917	1.3019	0.0470	**	0.3943
10^5	1.2341	0.0574	**	0.6287	1.2618	0.0100	**	0.9953
10^6	1.2654	0.0474	**	0.9405	1.2609	0.0076	**	0.8995

The estimation of the correlation dimension using different kernel functions can vary. The estimation of the dimension for the Sierpinski carpet for different kernel functions is presented in Tab. 10.5. The table shows the theoretical dimension D_2 and its estimate $\widehat{D_2}$ together with the standard deviation sd. The recommended range for L_4 regression is also included, where $f_{\min} = \log \Omega_{\min}$ and $f_{\max} = \log \Omega_{\max}$.

Table 10.5: Sierpinski carpet analysis for the different kernel functions.

Kernel function	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
H_2	1.8928	1.9851	0.2625	**	0.7251	1.0	3.0
H_4	1.8928	1.8673	0.1128	**	0.8212	1.0	3.0
H_7	1.8928	1.9148	0.0863	**	0.7988	1.0	3.0
H_{10}	1.8928	1.9019	0.0636	**	0.8862	1.0	3.0
H_∞	1.8928	1.8958	0.0559	**	0.9572	1.0	3.0

With the increasing dimension of the kernel function H_n, the standard deviation decreases. The estimates were unbiased in all cases, however, the most accurate estimation occurred for H_∞. The recommended intervals for the regression were

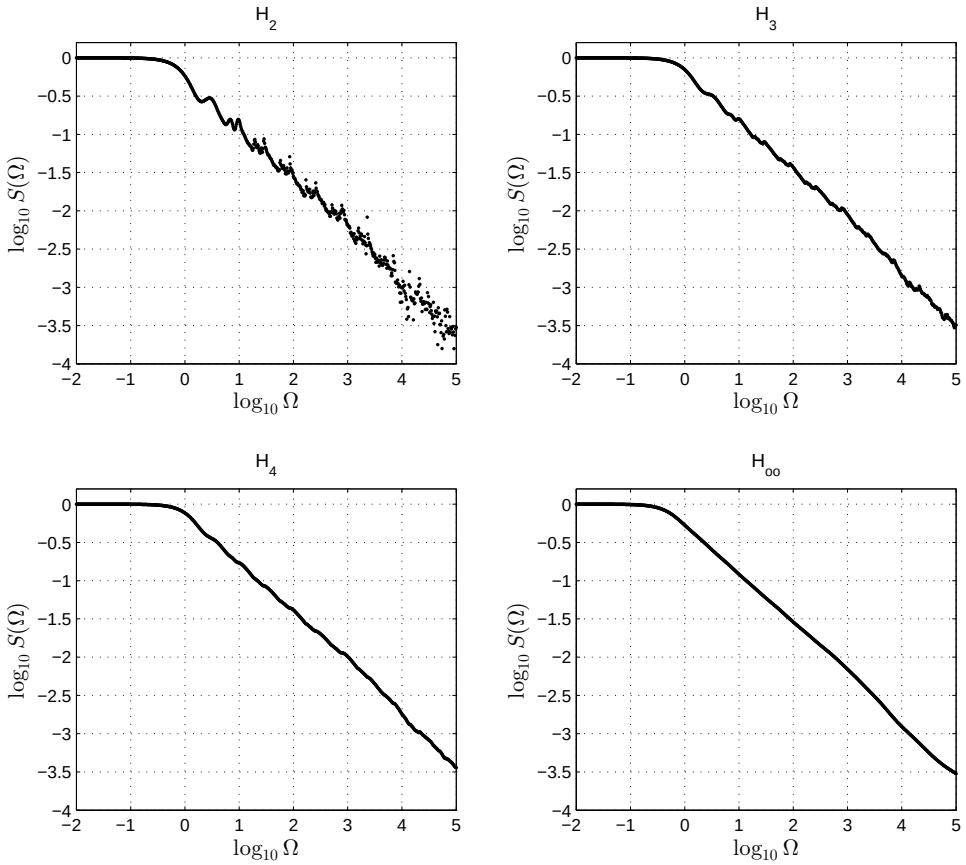


Figure 10.2: Rotational spectra of the Sierpinski carpet.

the same in all cases, which means that they were independent of the kernel function; nevertheless, as will be seen later, they will be dependent on the theoretical dimension of the fractal structure.

We will verify that the rotational spectrum provides unbiased estimate of the correlation dimension. The following experiments are conducted for parameters $M = 10^5$ and kernel function H_∞ for various fractal sets. The parameter n denotes the topological dimension of the space, where the set is investigated. The analysis of n-dimensional Cantor dust is presented in Tab 10.6, the Takagi function is analysed in Tab 10.7, the fBm trajectory correlation dimension is presented in Tab 10.8, the Levy flight results are in Tab. 10.9 and finally the results of n-dimensional fBm trajectory are summarized in Tab. 10.10.

The results presents unbiased estimation of correlation dimension for a variety of sets in one or multiple dimensions. The variance of the estimate decreases with increasing number of points, though the minimal recommended sample size is $M \geqslant 10^5$ for infinite-dimensional kernel H_∞. The suggested frequency range $f_{\min}$ and $f_{\max}$ for regression is still dependent on the fractal set type, though in general the optimal thresholds are higher for fractal sets with higher theoretical correlation dimension.

Table 10.6: Correlation dimension of Cantor dust.

n	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
1	0.6309	0.6324	0.0053	**	0.7772		
2	1.2619	1.2765	0.0136	**	0.2830	0.0	2.0
3	1.8928	1.9065	0.0148	**	0.3278		

Table 10.7: Correlation dimension of Takagi function.

n	a	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
	0.55	1.1375	1.1108	0.0324	**	0.4099	0.5	1.5
	0.60	1.2630	1.2469	0.0179	**	0.3684	1.0	2.0
	0.65	1.3785	1.3547	0.0801	**	0.7664	0.5	1.5
	0.70	1.4854	1.4981	0.0577	**	0.8258	1.0	2.0
2	0.75	1.5850	1.5889	0.0169	**	0.8175	1.0	1.9
	0.80	1.6780	1.6961	0.0401	**	0.6517	0.9	1.9
	0.85	1.7655	1.7731	0.0141	**	0.5899	1.1	1.9
	0.90	1.8480	1.8475	0.0149	**	0.9732	1.1	1.8
	0.95	1.9260	1.9206	0.0096	**	0.5738	1.1	1.7

Table 10.8: Correlation dimension of fBm trajectory.

n	H	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
	0.1	1.9	1.9152	0.0233	**	0.5142	0.9	1.5
	0.2	1.8	1.8193	0.0113	*	0.0876	0.9	1.5
	0.3	1.7	1.7003	0.0136	**	0.9824	0.7	1.5
	0.4	1.6	1.6068	0.0168	**	0.6857	1.0	1.5
1	0.5	1.5	1.4946	0.0075	**	0.4715	1.0	2.0
	0.6	1.4	1.4218	0.0392	**	0.5781	1.0	2.0
	0.7	1.3	1.2964	0.0981	**	0.9707	0.7	2.0
	0.8	1.2	1.1677	0.0512	**	0.5281	0.8	2.0
	0.9	1.1	1.1019	0.0070	**	0.7861	0.9	1.5

Table 10.9: Correlation dimension of two-dimensional Levy flight.

n	a	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
2	0.2	0.2	0.1931	0.0235	**	0.7691	−16.5	−14.5
	0.3	0.3	0.2948	0.0134	**	0.6980	−14.0	−12.0
	0.4	0.4	0.4036	0.0163	**	0.8252	−10.0	−8.0
	0.5	0.5	0.4946	0.1098	**	0.9608	−8.0	−6.0
	0.6	0.6	0.6048	0.0184	**	0.7942	−5.0	−3.0
	0.7	0.7	0.7252	0.0372	**	0.4981	−4.0	−2.0
	0.8	0.8	0.8013	0.0718	**	0.9856	−5.0	−2.0
	0.9	0.9	0.9343	0.0944	**	0.7163	−3.0	−1.0
	1.0	1.0	0.9711	0.0621	**	0.6417	−3.0	0.0

Table 10.10: Correlation dimension of n-dimensional fBm trajectory.

n	H	D_2	$\widehat{D_2}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
3	0.75	1.3333	1.3334	0.0302	**	0.9974	1.5	2.0
3	0.50	2.0000	2.0083	0.0203	**	0.6826	1.3	1.8
3	0.33	3.0000	2.9857	0.0580	**	0.8053	1.2	1.7
4	0.75	1.3333	1.3328	0.0165	**	0.9758	1.4	1.9
4	0.5	2.0000	1.9958	0.0203	**	0.8361	1.3	1.8
4	0.33	3.0303	3.0503	0.0467	**	0.6685	1.0	1.5
4	0.25	4.0000	4.0770	0.0959	**	0.4220	0.9	1.4

10.2 Estimating D_α Using Modified Renyi Entropy

The methodology of estimating Renyi dimension using the modified entropy approach employs its dependence on the logarithm of ϵ as

$$D_\alpha^* = \lim_{\epsilon \to 0^+} \frac{H_\alpha^*(\epsilon)}{-\ln \epsilon} \tag{10.4}$$

and can be estimated using model with additional Gaussian noise $e \sim \mathcal{N}(0; \sigma^2)$ in the form

$$H_\alpha^* = A - D_\alpha^* \ln \epsilon + e. \tag{10.5}$$

We can use least squares method for D_α^* estimation using different values ϵ_i for $i = 1, \ldots, N$. We suggest to use geometrically increasing series of ϵ_i generated by formula

$$\epsilon_i = 10^{f_{\min} + (i-1)\Delta f} \tag{10.6}$$

with $N = \lfloor (f_{\max} - f_{\min})/\Delta f \rfloor + 1$. The entropy dependency on parameter ϵ is visualized in Fig. 10.3.

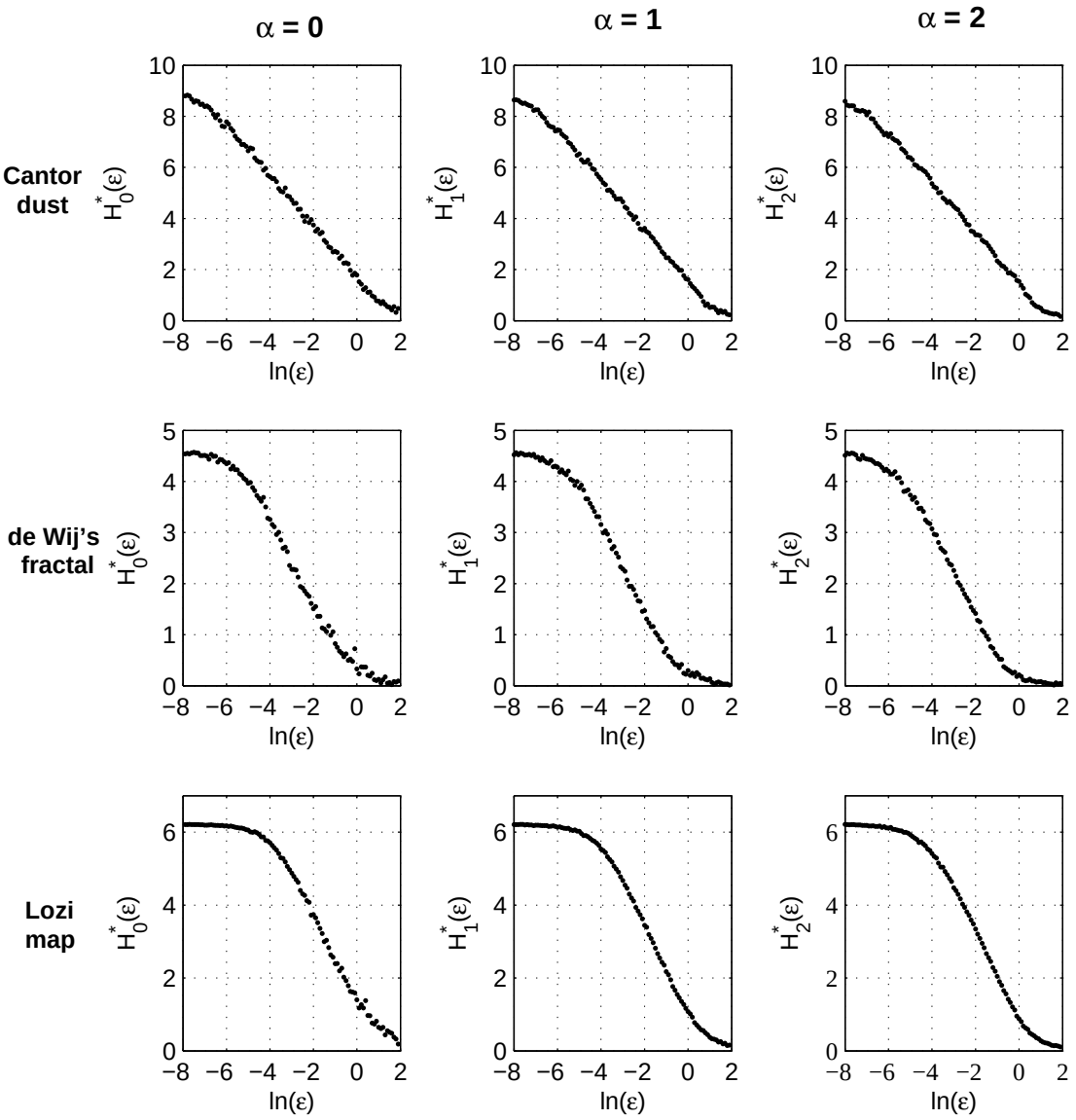


Figure 10.3: Renyi entropy for various α and Cantor dust (top), de Wij's fractal (middle) and Lozi map (bottom).

At first, we would like to prove that the estimate is unbiased with increasing number of data points. We have theoretically deduced that both correlation and capacity dimensions can be estimated using the modified entropy. We will use the example of Cantor set to perform the estimation of dimension for various number of input points. The results for capacity dimension are in Tab. 10.11, while the results for correlation dimension are displayed in Tab. 10.12. As easy to see, the standard deviation decreases with increasing number of points, while keeping its estimation unbiased.

Table 10.11: Capacity dimension estimation of Cantor set.

N	D_2	$\widehat{D_2^*}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
10^3	0.6309	0.6295	0.0343	**	0.9674	-3.0	-1.0
10^4	0.6309	0.6305	0.0239	**	0.9866	-3.0	-1.0
10^5	0.6309	0.6327	0.0034	**	0.5965	-3.0	-1.0
10^6	0.6309	0.6311	0.0018	**	0.9115	-3.0	-1.0

Table 10.12: Correlation dimension estimation of Cantor set.

N	D_2	$\widehat{D_2^*}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
10^3	0.6309	0.6215	0.0465	**	0.8398	-3.0	-1.0
10^4	0.6309	0.6344	0.0134	**	0.7939	-3.0	-1.0
10^5	0.6309	0.6286	0.0047	**	0.6246	-3.0	-1.0
10^6	0.6309	0.6301	0.0015	**	0.5938	-3.0	-1.0

As next, we would like to demonstrate the unbiasedness of the method using the comparison with traditional box-counting approach and model (3.21). For a fair comparison we have chosen the dimension estimation of discrete dynamic systems. Discrete dynamic systems with chaotic behaviour generate fractal trajectories and attractors with nonlinear character. The investigation of this kind of sets can be performed in to ways – in the original state space or in the reconstructed space. Recall that any Renyi dimension D_α of reconstructed attractor is the same as in the case of state space.

Tab.10.13 is showing the comparison of dimension estimation using box-counting (denoted as box-count) method and new method of modified Renyi entropy (denoted as m. Renyi). For the experiment, Henon map with parameters $a = 0.4$, $b = 0.3$ and starting points $x_{0,1} = 0, x_{0,2} = 0.9$ and Lozi map with parameters $a = 1.7$, $b = 0.5$ and starting points $x_{0,1} = -0.1, x_{0,2} = 0.1$ were used for the simulation for $M = 10^5$ points. The analysis was studied in two cases, in the case of resulting sample of the discrete dynamical system (point set) and in the case of state space reconstruction (reconstr.). The estimates in the case of the point set were not biased in both

Table 10.13: Discrete dynamical system analysis.

system	α	D_α	data	method	$\widehat{D_\alpha}$	sd	z	p-value
Henon map	1	1.2583	point set	m. Renyi	1.2608	0.0156	**	**0.8727**
				box-count	1.2428	0.0113	*	**0.1702**
			reconst.	m. Renyi	1.2590	0.0056	**	**0.9005**
				box-count	1.2489	0.0031		0.0024
Henon map	2	1.2201	point set	m. Renyi	1.2243	0.0174	**	**0.8093**
				box-count	1.2161	0.0109	**	**0.7136**
			reconst.	m. Renyi	1.2230	0.0026	**	**0.2647**
				box-count	1.2172	0.0014		0.0383
Lozi map	1	1.4042	point set	m. Renyi	1.4131	0.0197	**	**0.6514**
				box-count	1.3915	0.0174	**	**0.4655**
			reconst.	m. Renyi	1.4098	0.0044	*	**0.2031**
				box-count	1.3945	0.0032		0.0024
Lozi map	2	1.3845	point set	m. Renyi	1.3937	0.0144	**	**0.5229**
				box-count	1.3786	0.0161	**	**0.7140**
			reconst.	m. Renyi	1.3885	0.0031	*	**0.1969**
				box-count	1.3749	0.0041		0.0192

approaches, however the reconstruction for W = 5 and the dimension estimation in 5-dimensional space has shown that both methods can achieve smaller standard deviations, but the resulting estimate was unbiased only in the case of modified Renyi entropy approach, as seen from the corresponding p-values. The experiment was also conducted for bigger lengths of the reconstruction sliding window, but it didn't have significant impact on the results and their precision.

When the systems are investigated in the state space of low dimension ($n = 2$), the box-counting offered more accurate estimates with smaller standard deviation than the novel method. However, the p-values indicate unbiasedness in both cases. Another behaviour of estimation methods has been observed in the case of state reconstruction when the space dimension is large ($n = 5$). Therefore, the box-counting estimates of event probabilities are biased due to data sparsity. Moreover, all the box-counting estimates after reconstruction are biased. The sparsity effect is not present in the case of new method, where the p-values are higher with similar standard deviation. Therefore, the modified Renyi dimension is more suitable for reconstructed systems in higher-dimensional space.

The following calculations are performed for $M = 10^5$ for respective boundaries $f_{\min}$ and $f_{\max}$ and various parameters α. At first, the calculation is performed for monofractals - self-similar sets that fulfil the open set condition, therefore their Hausdorff dimension equals Renyi dimension for any eligible parameter α, e.g. $D_{\mathrm{H}} = D_0 = D_2$. The results of capacity dimension estimation are provided in Tab. 10.14 and the estimates in the case of correlation dimension are in the Tab. 10.15 for $\Delta f = 0.05$ and $M = 10^5$. The theoretical capacity (correlation) dimension is denoted D_0

(D_2) whereas its estimate is $\widehat{D_0^*}$ ($\widehat{D_2^*}$) together with its standard deviation sd. The range for the choice of $\ln \epsilon$ is recommended to be in interval $[f_{\min}; f_{\max}]$.

Table 10.14: Capacity dimension analysis.

System	a	D_0	$\widehat{D_0^*}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
Cantor set	1/4	0.5000	0.5059	0.0070	**	0.3993	-3.0	-1.0
Cantor set	1/3	0.6309	0.6327	0.0034	**	0.5965	-3.0	-1.0
Cantor dust	1/4	1.0000	0.9834	0.0157	**	0.2937	-2.0	0.0
Cantor dust	1/3	1.2619	1.2547	0.0133	**	0.5883	-2.0	0.0
Even numbers set	-	0.6990	0.7030	0.0148	**	0.7870	-4.0	-1.0
Sierpinki carpet	1/3	1.8928	1.8894	0.0059	**	0.5644	-2.0	-1.0
Sierpinki carpet	1/4	1.5000	1.4901	0.0148	**	0.5035	-2.0	-1.0

Table 10.15: Correlation dimension analysis.

System	a	D_2	$\widehat{D_2^*}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
Cantor set	1/4	0.5000	0.4974	0.0034	**	0.4444	-3.0	-1.0
Cantor set	1/3	0.6309	0.6286	0.0047	**	0.6246	-3.0	-1.0
Cantor dust	1/4	1.0000	0.9863	0.0221	**	0.5353	-2.0	-1.0
Cantor dust	1/3	1.2619	1.2630	0.0269	**	0.9674	-2.0	0.0
Even numbers set	-	0.6990	0.6991	0.0038	**	0.9790	-4.0	-1.0
Sierpinki carpet	1/3	1.8928	1.8964	0.0083	**	0.6645	-2.0	-1.0
Sierpinki carpet	1/4	1.5000	1.5053	0.0064	**	0.4076	-2.0	-1.0

The graph of De Wijs fractal with parameter a is a kind of multifractal that has Renyi dimension dependent on the dimension parameter α. Recall its corresponding Renyi dimension equals

$$D_\alpha = \frac{1}{1-\alpha} \log_2 \left(a^\alpha + (1-a)^\alpha\right) \tag{10.7}$$

for $0 < a < 1/2$ and $\alpha \in [0; 1) \cup (1, \infty)$ with particular case

$$D_1 = \lim_{\alpha \to 1} D_\alpha = -a \log_2 a - (1-a) \log_2 (1-a). \tag{10.8}$$

The D_α^* has been estimated for $\alpha \in \{0, 1/2, 1, 3/2, 2\}$ and the testing results are included in Tab. 10.16. In all cases we accepted the hypothesis that the estimated dimension is unbiased estimation of the true Renyi dimension of the fractal set.

The experiments exhibited that the modified Renyi dimension can be suitable for dimension estimation both in time space and reconstructed space. The minimal recommended number of points for analysis is $M \geqslant 10^5$. We have verified that the

method works reliably for capacity, information and correlation dimension as well as fractional dimensions with $\alpha \in \{1/2, 3/2\}$. Numerical calculations further showed that recommended thresholds for $f_{\min}$ and $f_{\max}$ can be obtained by focusing only on the dependent value of modified Renyi entropy. For any set with M points, the maximum value of modified Renyi entropy is $\ln M$. We recommend to use values ϵ that satisfy $H^*_\alpha(\epsilon) \geqslant \ln M/2$ and $H^*_\alpha(\epsilon) \leqslant 0.7 \ln M$.

Table 10.16: De Wijs fractal dimensions.

α	a	D_α	$\widehat{D^*_\alpha}$	sd	z	p-value	$f_{\min}$	$f_{\max}$
0	$a = 1/3$	1.0000	0.9908	0.0058	*	0.1127	-6.0	-4.0
0	$a = 1/4$	1.0000	0.9890	0.0087	*	0.2062	-6.0	-4.0
0	$a = 1/6$	1.0000	0.9780	0.0143	*	0.1240	-6.0	-4.0
1/2	$a = 1/3$	0.9581	0.9574	0.0062	**	0.9101	-5.5	-3.5
1/2	$a = 1/4$	0.9000	0.8921	0.0103	**	0.4431	-5.5	-3.5
1/2	$a = 1/6$	0.8035	0.7895	0.0159	**	0.3786	-5.0	-3.0
1	$a = 1/3$	0.9183	0.9158	0.0060	**	0.6769	-4.0	-2.0
1	$a = 1/4$	0.8250	0.8259	0.0098	**	0.9269	-4.0	-2.0
1	$a = 1/6$	0.6500	0.6387	0.0217	**	0.6026	-3.0	-1.0
3/2	$a = 1/3$	0.8814	0.8749	0.0099	**	0.5115	-4.0	-2.0
3/2	$a = 1/4$	0.7376	0.7255	0.0153	**	0.4290	-4.0	-2.0
3/2	$a = 1/6$	0.5419	0.5234	0.0209	**	0.3761	-3.0	-1.0
2	$a = 1/3$	0.8480	0.8359	0.0189	**	0.5220	-3.0	-1.0
2	$a = 1/4$	0.6781	0.6687	0.0205	**	0.6466	-3.0	-1.0
2	$a = 1/6$	0.4695	0.4552	0.0235	**	0.5429	-2.0	0.0

10.3 Estimating H from Short Time Series

This section experimentally verifies the method of dfBB parameter estimation on short time series. To be able to verify the deduction of the dfBB correlation function, it is necessary to apply precise fGn sample generator with correlation function specified by formula

$$\eta(m, H) = \frac{1}{2}\left(|m+1|^{2H} - 2|m|^{2H} + |m-1|^{2H}\right) \tag{10.9}$$

for $m = 0, \ldots, N$. At first, we validate that it is possible to simulate the fGn with proper autocorrelation structure, and on the basis of fGn sample one can reconstruct the dfBB and test its properties. Therefore, it is needed to investigate the quality of the generator by means of testing the accuracy of correlation coefficients for different Hurst exponents H. Four generators of fGn were tested – Circulant embedding method (CEM), Lowen method (LM), Abry-Sellan wavelet method (ASM) and Randomly Stimulated Method (RSM) that uses FIR filter [81, 82]. For Q realisations

of fGn of length $M = 10^6$ we denote the estimated i-th correlation coefficient from j-th realisation as ρ_i^j. For given H, the theoretical i-th correlation coefficient is denoted as $\rho_i^{\mathrm{TH}} = \eta(i, H)$. The null hypothesis for one-sample t-test states that the expected value of k-th correlation coefficient equals the theoretical value

$$H_{0(k)} : \mathrm{E}(\rho_k) = \rho_k^{\mathrm{TH}}. \tag{10.10}$$

The p-values for ρ_1 are shown in Tab. 10.17, whereas the Tab. 10.18 displays the p-values for second autocorrelation coefficient ρ_2. The symbol • indicates that the p-value is smaller than 10^{-10}. The estimates of ρ_1, ρ_2 are unbiased when p-value $> \alpha$ for critical level $\alpha = 0.05$. As seen, the CEM offers unbiased estimates for $H \leqslant 0.75$. The LM is unbiased only for $H \leqslant 0.5$ and ASM just for $H = 1/2$. Considering the expected values $\mathrm{E}(\rho_1)$ and $\mathrm{E}(\rho_2)$ as the descriptors of the problem, only RSM provided unbiased correlation coefficients in the whole range of H. The RSM method employs FIR filter [83] of length $L = 10$ that was designed to generate signal from input iid noise from standard normal distribution. The coefficients a_i $(i = 1, \ldots, N)$ of FIR filter were calculated to satisfy

$$r_n = \frac{1}{N - n - 1} \sum_{j=0}^{N-n-1} a_n a_{j+n} \tag{10.11}$$

for $n = 0, \ldots, L - 1$, where r_n is n-th autocovariance coefficient of fGn.

Table 10.17: Unbiasedness of ρ_1 coefficient of dfBB as p-value from (10.10).

method	H									
	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.98	0.99
CEM	0.93	0.78	0.32	0.35	0.97	0.76	0.32	•	•	•
LM	0.26	0.82	0.16	0.19	0.55	•	•	•	•	•
ASM	•	•	•	•	0.66	$4 \cdot 10^{-7}$	•	•	•	•
RSM	0.38	0.69	0.66	0.66	0.38	0.45	0.50	0.33	0.82	0.39

Table 10.18: Unbiasedness of ρ_2 coefficient of dfBB as p-value from (10.10).

method	H									
	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.98	0.99
CEM	0.68	0.25	0.16	0.35	0.40	0.12	0.01	•	•	•
LM	0.98	0.42	0.15	0.33	0.40	•	•	•	•	•
ASM	•	•	•	$2 \cdot 10^{-7}$	0.73	$3 \cdot 10^{-6}$	•	•	•	•
RSM	0.30	0.62	0.34	0.89	0.91	0.28	0.51	0.68	0.85	0.64

It was discovered that not all fGn generators produce samples with sufficient statistical properties so that they would be usable for further testing. Therefore, we investigate only samples generated by RSM and CEM, that were proven to provide satisfactory samples. Recall that in this method, we first generate Q independent samples of fGn, which are summarized and transferred into dfBB sample. At first, we would like to prove the dependency of the estimate $\widehat{H}$ and its standard deviation sd on the number of generated samples Q for theoretical value of $H = 0.4$ and sample length $N = 10$. Tab. 10.19 summarizes the results for both CEM and RSM methods and illustrates the decay of the standard deviation in both cases.

Table 10.19: Analysis of dfBB for $H = 0.4$.

Q	RSM				CEM			
	$\widehat{H}$	sd	z	p-value	$\widehat{H}$	sd	z	p-value
1	0.3183	0.6524	**	0.9003	0.2541	0.7811	**	0.8518
10	0.3610	0.1869	**	0.8347	0.4991	0.2439	**	0.6845
100	0.4563	0.0617	**	0.3615	0.4528	0.0905	**	0.5596

For the following calculations we allow $Q = 100$, therefore the estimation has been done from $N \cdot Q = 1000$ data points in total. The results and corresponding p-values are included in Tab. 10.20 for RSM and Tab. 10.21 for CEM. As far as RSM generator is concerned, the methodology provides fair and unbiased estimate of Hurst exponent in the whole range of H. Due to the inaccuracy of CEM generation for higher values of H, we get the unbiased estimate of Hurst exponent only for $H < 0.8$.

Table 10.20: Hurst exponent estimate using RSM for $N = 10$ (left) and $N = 20$ (right).

H	$\widehat{H}$	sd	z	p-value
0.1	0.0806	0.0872	**	0.8289
0.2	0.2316	0.0792	**	0.6992
0.3	0.3454	0.0909	**	0.6175
0.4	0.4563	0.0617	**	0.3853
0.5	0.4619	0.0572	**	0.5054
0.6	0.6510	0.0834	**	0.5409
0.7	0.6848	0.0933	**	0.8706
0.8	0.7544	0.0570	**	0.4237
0.9	0.9617	0.0703	**	0.3801

H	$\widehat{H}$	sd	z	p-value
0.1	0.1081	0.0848	**	0.9239
0.2	0.2244	0.0579	**	0.6735
0.3	0.2991	0.0458	**	0.9843
0.4	0.3682	0.0526	**	0.5455
0.5	0.5428	0.0792	**	0.5889
0.6	0.6203	0.0478	**	0.6711
0.7	0.6905	0.0535	**	0.8591
0.8	0.8441	0.0482	**	0.3602
0.9	0.9528	0.0681	**	0.4381

The standard deviation of Hurst exponent estimate performed at samples of length $N = 20$ was in the majority of cases lower than in the case $N = 10$. The increasing sample length can reduce the uncertainty of estimate and the method can be reliably used for any length of the time series, but the recommended number of input samples is (at least) $Q = 100$. The main advantages of this method is the ability to process short independent samples in the whole range of possible Hurst exponent. To the

Table 10.21: Hurst exponent estimate using CEM for $N = 10$ (left) and $N = 20$ (right).

H	$\widehat{H}$	sd	z	p-value
0.1	0.1486	0.1299	**	0.7083
0.2	0.1700	0.1383	**	0.8283
0.3	0.3103	0.0989	**	0.9171
0.4	0.4528	0.0905	**	0.5596
0.5	0.4627	0.0880	**	0.6717
0.6	0.5749	0.1010	**	0.8037
0.7	0.6156	0.1183	**	0.4756
0.8	0.6467	0.0646		0.0176
0.9	0.7227	0.0447		0.0001

H	$\widehat{H}$	sd	z	p-value
0.1	0.0870	0.1112	**	0.9069
0.2	0.1971	0.0666	**	0.9653
0.3	0.3371	0.1039	**	0.7210
0.4	0.4328	0.0771	**	0.6705
0.5	0.5255	0.0714	**	0.7210
0.6	0.5687	0.0424	**	0.4604
0.7	0.7162	0.0443	**	0.7146
0.8	0.7059	0.0675	*	0.1633
0.9	0.7454	0.0714		0.0304

disadvantages belongs mainly the focusing on the several first correlation coefficients which cannot discover the dependencies between more distant time series elements and bigger standard deviation of the estimate. This technique is recommended in two cases:

- analysis of short time series with at least 10 elements,
- analysis of stationary time series using signal segmentation which satisfies segment independence.

The second approach leads to more accurate estimate of H which we later demonstrate experimentally on the signals from biological autoluminiscence.

10.4 Hurst Exponent Estimation in 1D Using Wishart Distribution

The experimental study of Hurst exponent estimation via Wishart distribution is at first conducted in one dimension on simulated data. There are a few well-known generators of one dimensional fBm, such as Lowen, Davies-Harte, Circulant embedding and Abry-Sellan methods. The generators can be used in two ways:

- produce fBm samples and estimate its parameters,
- produce fBm samples and transform them into another series (e.g. fBB), which autocorrelation function is also known.

The parameter estimate can be obtained via solving optimization task (9.14) and the confidence regions for H and σ can be calculated using equation (9.17). In the following experiments we will be using standardized fBm, therefore its variance is normalized to $\sigma = 1$ for various values of Hurst exponent H.

At first, the confidence regions (9.17) were investigated for $H \in (0, 1)$ and fBm, fGn, fBB and dfBB models. Experiments showed that the confidence regions using fractional motions and noises models (fBm and fGn) have the same shape and area for the whole spectrum of parameter H. The confidence region for fractional bridges (fBB and dfBB) are wider, but again alike within this process family. The confidence regions for theoretical value $H = 0.25$ generated by the Lowen method and fBm model for different sample sizes is depicted in Fig. 10.4, whereas the confidence interval using fBB method for the same H is in Fig. 10.5. The circle symbol marks the theoretical value of $(H, \ln \sigma)$, whereas the star denotes its estimate. With the increasing length of the generated time series, the confidence region is getting more narrow, however, still containing the theoretical value of H. For the rest of this chapter, we will not differentiate between fBm and fGn models, neither between fBB and dfBB since the results coming from these approaches are almost alike. In the rest of the calculations, the fBm and fBB models are used for illustration of the method performance. The standardized fBm was examined for different input sample sizes n and the Hurst exponent estimates and their standard deviation is illustrated in Tab. 10.22. As expected, the standard deviation of the estimate decreases with increasing sample size and the estimate converges to the theoretical value.

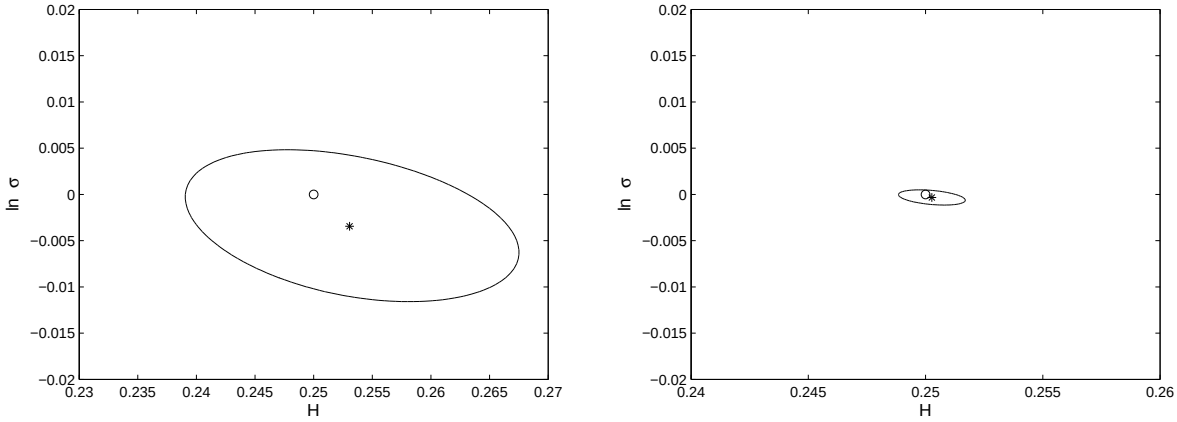


Figure 10.4: Confidence region of fBm model with $N = 10$ for $n = 10^4$ (left) and $n = 10^6$ (right) generated by Lowen method.

Table 10.22: Analysis of fBm and fBB for $H = 0.25$.

n	fBm				fBB			
	$\widehat{H}$	sd	z	p-value	$\widehat{H}$	sd	z	p-value
10^3	0.25174	0.022340	**	0.9379	0.25284	0.032865	**	0.9311
10^4	0.25071	0.010592	**	0.9466	0.25079	0.014354	**	0.9561
10^5	0.25083	0.002175	**	0.7028	0.25012	0.003193	**	0.9700
10^6	0.24980	0.001078	**	0.8528	0.25069	0.001436	**	0.6309

The next step is verification of the generation methods in one dimension for both fBm and fBB models. At first, we estimated the Hurst exponent of a time series generated

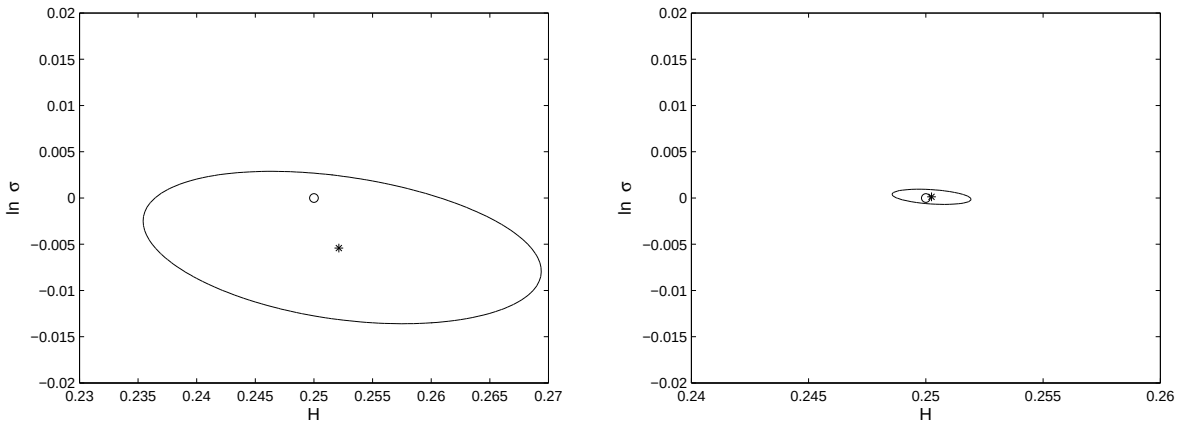


Figure 10.5: Confidence region of fBB model with $N = 10$ for $n = 10^4$ (left) and $n = 10^6$ (right) generated by Lowen method.

by the Lowen method. The time-series of length $2 \times (N+1) \times n$ was always generated and divided into segments of size $N + 1$. For the analysis, only even segments were taken into consideration, due to the memory of fBm and fBB processes. Removing odd segments from the initial sample assures there is no correlation between the segments and their independence. The total number of segments for analysis was $n = 10^5$ and the study has been performed for $N = 10$. Therefore, the parameter of Wishart distribution is $p = 10$ for fBm and fGn models, but $p = 9$ for fBB and dfBB. The results for Hurst exponent estimation are in Tab. 10.23, while H denotes the theoretical value of Hurst exponent, $\widehat{H}$ is its estimate, *sd* is standard deviation. Column z denotes the unbiasedness of the estimate, having one-star symbol (*) when z-score is lower than 1.96 and two stars (**) if z-score is lower than 1.24. The standard deviation of H estimate was always smaller when using the fBm method in contrast to the fBB method, however, for all $H \geqslant 0.1$, the estimate was unbiased.

Table 10.23: Estimation of H from short samples ($N = 10, n = 10^5$) generated by Lowen method.

H	fBm				fBB			
	$\widehat{H}$	*sd*	z	p-value	$\widehat{H}$	*sd*	z	p-value
0.05	0.07990	0.001631		$< 10^{-5}$	0.06923	0.002541		$< 10^{-5}$
0.10	0.09880	0.001716	**	0.4844	0.10065	0.002698	**	0.8096
0.15	0.15199	0.001935	**	0.3038	0.14781	0.002906	**	0.4511
0.20	0.19995	0.002076	**	0.9808	0.19995	0.003048	**	0.9869
0.25	0.25083	0.002175	**	0.7028	0.25012	0.003193	**	0.9700
0.30	0.30060	0.002268	**	0.7914	0.30027	0.003303	**	0.9349
0.35	0.34902	0.002289	**	0.6686	0.34878	0.003397	**	0.7195
0.40	0.39999	0.002323	**	0.9966	0.39878	0.003440	**	0.7229
0.45	0.45057	0.002306	**	0.8048	0.44998	0.003533	**	0.9955
0.50	0.50061	0.002283	**	0.7893	0.50275	0.003543	**	0.4376

Very similar results were obtained by the Davies–Harte method and Circulant Embedding. Hurst exponent generated from the whole spectra from 0.05 to 0.95 was estimated both using fBm and fBB methods. The fBm model again provided better performance and a smaller standard deviation for all H except $H = 0.95$, where the estimate was biased. Complete results are in Tab. 10.24 for Davies–Harte and Tab. 10.25 for circulant embedding.

Table 10.24: Estimation of H from short samples ($N = 10, n = 10^5$) generated by Davies-Harte method.

H	fBm				fBB			
	$\widehat{H}$	sd	z	p-value	$\widehat{H}$	sd	z	p-value
0.05	0.05980	0.001187		$< 10^{-5}$	0.07451	0.001967		$< 10^{-5}$
0.10	0.09877	0.001435	**	0.3914	0.10210	0.001960	**	0.2840
0.15	0.15222	0.001652	*	0.1790	0.15442	0.002510	*	0.0782
0.20	0.19887	0.001801	**	0.5304	0.20451	0.003229	*	0.1625
0.25	0.25257	0.001947	*	0.1868	0.24976	0.003627	*	0.9472
0.30	0.30175	0.002138	**	0.4131	0.30795	0.004149	*	0.0553
0.35	0.35022	0.002047	**	0.9144	0.34893	0.004579	**	0.8152
0.40	0.39673	0.002098	*	0.1191	0.39597	0.003467	**	0.2451
0.45	0.44987	0.002128	**	0.9513	0.44899	0.003486	**	0.7720
0.50	0.50014	0.002121	**	0.9474	0.50280	0.003547	**	0.4299
0.55	0.55041	0.002132	**	0.8475	0.55340	0.003570	**	0.3409
0.60	0.60201	0.002083	**	0.3346	0.59880	0.003595	**	0.7385
0.65	0.64995	0.002033	**	0.9804	0.65293	0.003649	**	0.4220
0.70	0.69876	0.003593	**	0.7300	0.70022	0.003609	**	0.9514
0.75	0.75105	0.001855	**	0.5714	0.74892	0.003082	**	0.7260
0.80	0.79553	0.003558	**	0.2090	0.79553	0.003558	**	0.2090
0.85	0.84511	0.003502	*	0.1626	0.84511	0.003502	*	0.1626
0.90	0.90206	0.001245	*	0.0980	0.90083	0.003454	**	0.8101
0.95	0.95057	0.001434	**	0.6910	0.95057	0.002934	*	0.8460

The experiment was conducted as well using Abry–Sellan algorithm for parameters H from the long-memory spectrum and the H estimates are presented in Tab. 10.26. We have analysed criterion MSE as well. The decadic logarithm of MSE is plotted against H in in Fig. 10.6. It is clear that the MSe is always lower in case of fBm than using fGn approach. Except extreme values of H close to 0.1, the MSE exhibits stable values in its whole range.

To sum up, the investigation has shown that even for large samples, the method is unbiased for all H between 0.1 and 0.95, having low standard deviation with an order of magnitude minus three. With increasing length of input time series n, the confidence region size is significantly reduced. Moreover, the experiments showed that the mean square error of H estimate is inversely proportional to the number of segments which could be further used to achieve more accurate estimates. The comparison between the fBm and fBB models has shown the fBm model was proved to have smaller estimation error and its accuracy is most dependent on the product

Table 10.25: Estimation of H from short samples ($N = 10, n = 10^5$) generated by Circulant Embedding method.

H	fBm				fBB			
	$\widehat{H}$	sd	z	p-value	$\widehat{H}$	sd	z	p-value
0.05	0.06005	0.001994		$< 10^{-5}$	0.07367	0.002529		$< 10^{-5}$
0.10	0.10032	0.001894	**	0.8658	0.09517	0.002655	*	0.0689
0.15	0.15017	0.002162	**	0.9373	0.14985	0.002934	**	0.9592
0.20	0.19984	0.002110	**	0.9396	0.20031	0.002944	**	0.9161
0.25	0.24977	0.002275	**	0.9195	0.24526	0.003054	**	0.1206
0.30	0.30035	0.002268	**	0.8774	0.29964	0.003127	**	0.9083
0.35	0.34988	0.002317	**	0.9587	0.35028	0.003317	**	0.9327
0.40	0.39982	0.002195	**	0.9346	0.40007	0.003502	**	0.9841
0.45	0.45034	0.002161	**	0.8750	0.45496	0.003536	*	0.1607
0.50	0.49985	0.002105	**	0.9432	0.50425	0.003564	**	0.2331
0.55	0.55129	0.002060	**	0.5312	0.55010	0.003602	**	0.9779
0.60	0.59971	0.002000	**	0.8847	0.60323	0.003392	**	0.3410
0.65	0.64986	0.001852	**	0.9397	0.64967	0.003423	**	0.9232
0.70	0.70024	0.001569	**	0.8784	0.69990	0.003364	**	0.9763
0.75	0.75127	0.001329	**	0.3393	0.74503	0.003460	*	0.1509
0.80	0.79988	0.000972	**	0.9017	0.80491	0.003422	*	0.1513
0.85	0.85027	0.000668	**	0.6861	0.85052	0.003413	*	0.8789
0.90	0.90037	0.001098	**	0.7361	0.90461	0.003338	*	0.1673
0.95	0.94582	0.000868		$< 10^{-5}$	0.94549	0.003469	*	$< 10^{-5}$

Table 10.26: Estimation of H from short samples ($N = 10, n = 10^5$) generated by Abry-Sellan method.

H	fBm				fBB			
	$\widehat{H}$	sd	z	p-value	$\widehat{H}$	sd	z	p-value
0.50	0.50042	0.002278	**	0.8537	0.50046	0.0035246	**	0.8962
0.55	0.55054	0.002235	**	0.8091	0.55093	0.0035687	**	0.7944
0.60	0.59971	0.000691	**	0.6747	0.59881	0.0035774	**	0.7394
0.65	0.64954	0.002097	**	0.8264	0.65654	0.0035517	*	0.0656
0.70	0.70024	0.000618	**	0.6978	0.69741	0.0035767	**	0.4690
0.75	0.74961	0.001854	**	0.8334	0.74483	0.0035599	*	0.1464
0.80	0.79889	0.001655	**	0.5024	0.80139	0.0035584	**	0.6961
0.85	0.84895	0.001440	**	0.4659	0.85071	0.0035218	**	0.8402
0.90	0.89828	0.001136	*	0.1300	0.89827	0.0035014	**	0.6212
0.95	0.94739	0.000219		$< 10^{-5}$	0.95153	0.0032925	**	0.6422

of $N \times n$ rather than individually on n or N. The recommended model is the fBm model for at least $n = 10^5$ data points and segment size $N \geqslant 10$ which assures, the resulting correlating coefficients have high precision.

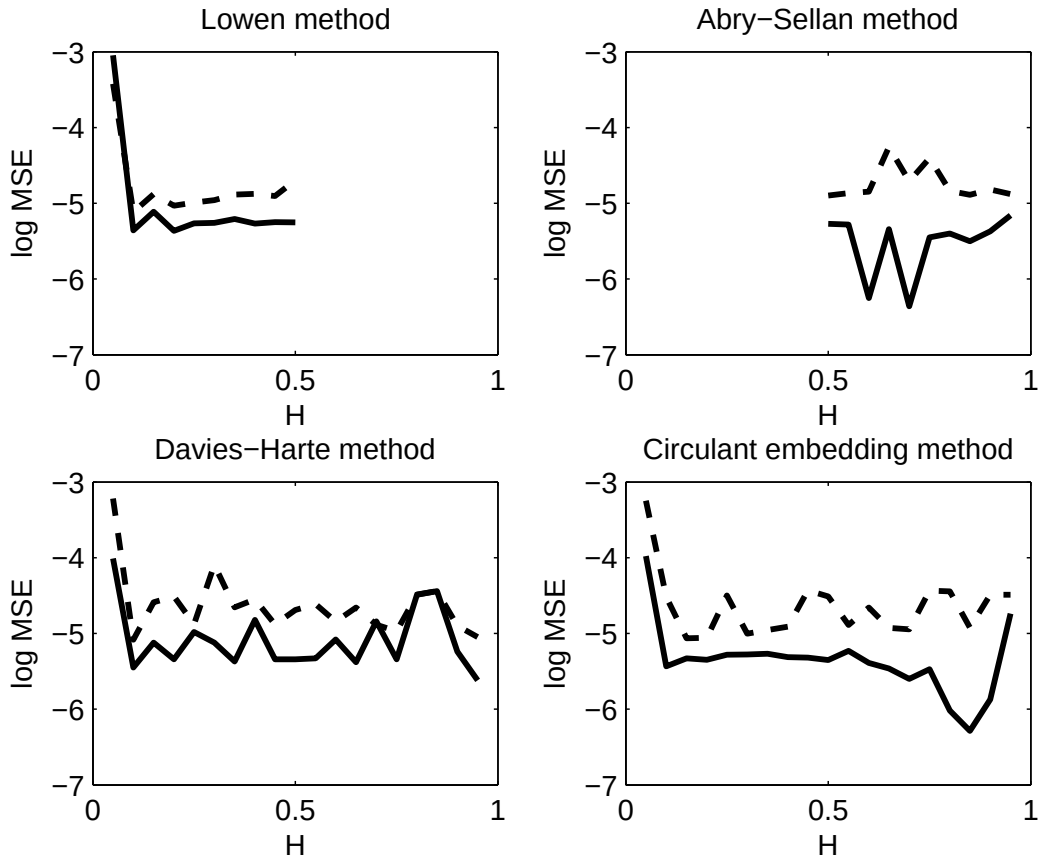


Figure 10.6: Decadic logarithm of mean square error (MSE) of H estimate for fBm (straight line) and fBB (dashed line).

10.5 Hurst exponent estimation in nD Using Wishart Distribution

Whereas in one dimension, there are a lot of methods that could be used for the simulation, in n-dimensional space the amount of methods is limited. In this section, the generalized circulant embedding method (GCE) is used for multidimensional fBm generation and the Hurst exponent is subsequently estimated via the Wishart distribution approach. Recall that thorough this chapter we are using slightly different notation than before, since $n \in \mathbb{N}$ denotes the dimension of the space where fBm is generated, $N \in \mathbb{N}$ still denotes the number of points in each segment sample and $m \in \mathbb{N}$ is the total number of samples of the n-dimensional fBm.

At first, we would like to validate the GCE generation method. Since the method can generate fBm in any dimension, we can easily validate it in one dimension and verify that it provides the same performance as traditional one-dimensional methods. The comparison between the traditional circulant embedding method (TCE) and generalized circulant embedding method is performed for the full range of parameter H and standardized fBm with unit variance. We investigated the optimal value of regularization parameter $\lambda > 0$ in one and two dimensions. It has

been observed the Wiener method of regularization numerically fails for $\lambda < 10^{-13}$ in many cases. Therefore, we suggest $\lambda = 10^{-12}$ as the minimum but safe value. Using this value of λ we can generate the n-dimensional fBm and estimate H using the maximum likelihood method and Wishart approach. The theoretical value of H is uniformly generated from 0.05 to 0.95.

The results for one dimensional case (traditional fractional Brownian motion) for TCE and GCE methods are shown in Tab. 10.27. While H denotes the theoretical value of Hurst exponent, $\widehat{H}$ is its estimate with corresponding standard deviation *sd*. We get unbiased estimate for all H except $H = 0.95$, when the fractional Brownian motion has a very long memory and both methods overestimate its theoretical value. Additionally, in all cases when the estimates are unbiased, the novel GCE method is having smaller MSE. The performance of TCE method is visualized in Fig. 10.7 as a 95% confidence interval of $H - \widehat{H}$.

Table 10.27: Hurst exponent estimation of fBm in one dimension for $N = 10$ and $m = 10^4$.

H	TCE				GCE			
	$\widehat{H}$	*sd*	z	p-value	$\widehat{H}$	*sd*	z	p-value
0.0500	0.0459	0.00276	*	0.1374	0.0498	0.00141	**	0.8872
0.1000	0.1010	0.00354	**	0.7776	0.0997	0.00180	**	0.8676
0.1500	0.1519	0.00405	**	0.6390	0.1538	0.00203	*	0.0612
0.2000	0.1923	0.00440	*	0.0801	0.2020	0.00222	**	0.3676
0.2500	0.2578	0.00472	*	0.0984	0.2479	0.00236	**	0.3736
0.3000	0.3013	0.00480	**	0.7865	0.3033	0.00246	*	0.1798
0.3500	0.3509	0.00502	**	0.8577	0.3487	0.00248	**	0.6001
0.4000	0.4068	0.00504	*	0.1773	0.4006	0.00252	**	0.8118
0.4500	0.4482	0.00493	**	0.7150	0.4494	0.00256	**	0.8147
0.5000	0.4956	0.00509	**	0.3873	0.5002	0.00250	**	0.9362
0.5500	0.5495	0.00497	**	0.9199	0.5512	0.00251	**	0.6326
0.6000	0.6001	0.00493	**	0.9838	0.6024	0.00249	**	0.3351
0.6500	0.6515	0.00477	**	0.7532	0.6501	0.00237	**	0.9663
0.7000	0.6942	0.00457	*	0.2044	0.7002	0.00222	**	0.9282
0.7500	0.7459	0.00440	**	0.3514	0.7490	0.00218	**	0.6464
0.8000	0.8020	0.00397	**	0.6144	0.7989	0.00197	**	0.5766
0.8500	0.8531	0.00346	**	0.3703	0.8490	0.00175	**	0.5677
0.9000	0.9052	0.00267	*	0.0515	0.8979	0.00147	*	0.1531
0.9500	0.9444	0.00175		0.0014	0.9594	0.00088		$< 10^{-5}$

In two dimensional space, where the equivalent of TCE does not exist, the experiment is performed for GCE and symmetric compact mask with $p = 1$ and $\rho = 4$, yielding $N = 80$. The estimation has been performed from $m = 10^4$ independent samples and the results are presented in Tab. 10.28. The method provided unbiased estimation of Hurst exponent for all $H < 0.90$, except the extreme case of long memory fractional Brownian surface. The performance of the GCE method is visualized in Fig. 10.8. Similar experiment was also conducted for three dimensional

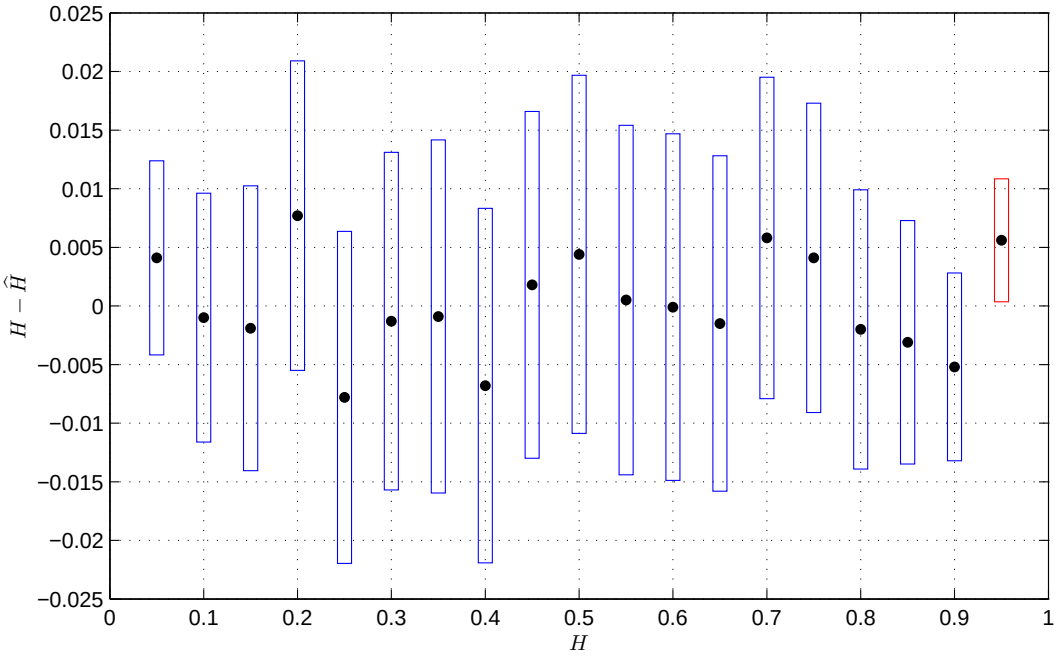


Figure 10.7: TCE method performance in one-dimensional space.

fBm where we used again symmetric compact mask with $p = 1$ and $\rho = 2$, yielding $N = 124$. The method was reliable for all cases where $H < 0.85$, whereas with long memory cases the estimates had low standard deviation, but were biased. The results from GCE method in three dimensions are contained in Tab. 10.29.

The results have shown that the GCE method can be used reliably for fBm generation except the cases with long memory. In that case the covariance function is having long dependence, where the differences between the autocorrelation coefficients are only tiny. Conducting further experiments, it was found that the performance of the estimation again depends on $m \cdot N$ rather than independently on m or N. When validating the unbiasedness of GCE, the experiments on fractional Brownian surface ($n = 2$) and fractional Brownian mass ($n = 3$) showed that the range of reliability for Hurst exponent shrinks its upper bound when the space dimension increases. For further analyses, it is recommended $N \cdot m \geqslant 10^5$ and for investigation of sets with $H < 0.95$ and $n = 1$, with $H < 0.9$ and $n = 2$ and with $H < 0.85$ and $n = 3$, respectively.

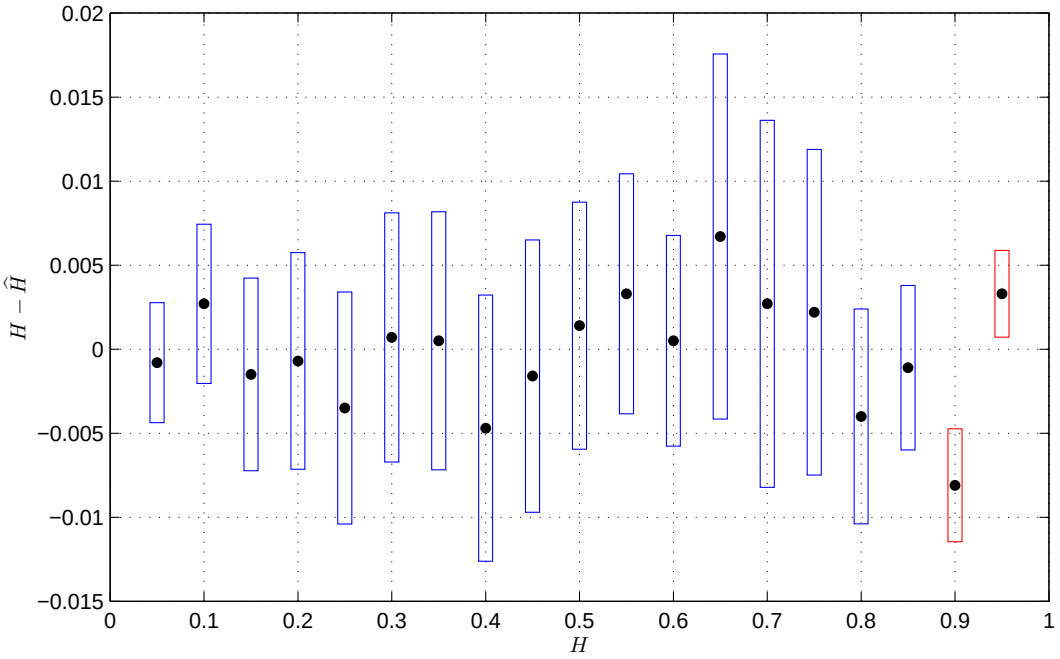


Figure 10.8: GCE method performance in two-dimensional space.

Table 10.28: Hurst exponent estimation of fBm in two dimensions for $N = 80$ and $m = 10^4$.

H	$\widehat{H}$	sd	z	p-value	MSE
0.0500	0.0508	0.00119	**	0.5014	$2.06 \cdot 10^{-6}$
0.1000	0.0973	0.00158	*	0.0875	$9.79 \cdot 10^{-6}$
0.1500	0.1515	0.00191	**	0.4323	$5.90 \cdot 10^{-6}$
0.2000	0.2007	0.00215	**	0.7447	$5.11 \cdot 10^{-6}$
0.2500	0.2535	0.00230	*	0.1281	$1.75 \cdot 10^{-5}$
0.3000	0.2993	0.00247	**	0.7769	$6.59 \cdot 10^{-6}$
0.3500	0.3495	0.00256	**	0.8451	$6.80 \cdot 10^{-6}$
0.4000	0.4047	0.00264	*	0.0750	$2.91 \cdot 10^{-5}$
0.4500	0.4516	0.00270	**	0.5535	$9.85 \cdot 10^{-6}$
0.5000	0.4986	0.00845	**	0.8684	$7.34 \cdot 10^{-5}$
0.5500	0.5467	0.00238	*	0.1656	$1.66 \cdot 10^{-5}$
0.6000	0.5995	0.00209	**	0.8109	$4.62 \cdot 10^{-6}$
0.6500	0.6433	0.00362	*	0.0642	$5.80 \cdot 10^{-5}$
0.7000	0.6973	0.00364	**	0.4582	$2.05 \cdot 10^{-5}$
0.7500	0.7478	0.00323	**	0.4958	$1.53 \cdot 10^{-5}$
0.8000	0.8040	0.00213	*	0.0604	$2.05 \cdot 10^{-5}$
0.8500	0.8511	0.00163	**	0.4998	$3.87 \cdot 10^{-6}$
0.9000	0.9081	0.00112		$< 10^{-5}$	$6.69 \cdot 10^{-5}$
0.9500	0.9467	0.00086		0.0001	$1.16 \cdot 10^{-5}$

Table 10.29: Hurst exponent estimation of fBm in three dimensions for $N = 124$ and $m = 10^3$.

H	$\widehat{H}$	sd	z	p-value	MSE
0.0500	0.0486	0.0051	**	0.7837	$2.75 \cdot 10^{-5}$
0.1000	0.0976	0.0072	**	0.7389	$5.70 \cdot 10^{-5}$
0.1500	0.1471	0.0089	**	0.7445	$8.74 \cdot 10^{-5}$
0.2000	0.1991	0.0053	**	0.8652	$2.88 \cdot 10^{-5}$
0.2500	0.2530	0.0019	*	0.1143	$1.27 \cdot 10^{-5}$
0.3000	0.3003	0.0024	**	0.9005	$6.07 \cdot 10^{-6}$
0.3500	0.3542	0.0058	**	0.4690	$5.15 \cdot 10^{-5}$
0.4000	0.4047	0.0078	**	0.5468	$8.31 \cdot 10^{-5}$
0.4500	0.4508	0.0077	**	0.9173	$5.93 \cdot 10^{-5}$
0.5000	0.4868	0.0095	*	0.1647	$2.64 \cdot 10^{-4}$
0.5500	0.5403	0.0064	*	0.1296	$1.35 \cdot 10^{-4}$
0.6000	0.5901	0.0060	*	0.0989	$1.34 \cdot 10^{-4}$
0.6500	0.6500	0.0052	**	0.9847	$2.73 \cdot 10^{-5}$
0.7000	0.6958	0.0042	**	0.3173	$3.60 \cdot 10^{-5}$
0.7500	0.7455	0.0029	*	0.1207	$2.84 \cdot 10^{-5}$
0.8000	0.7958	0.0024	*	0.0801	$2.32 \cdot 10^{-5}$
0.8500	0.7951	0.0018		$< 10^{-5}$	$3.02 \cdot 10^{-3}$
0.9000	0.8306	0.0015		$< 10^{-5}$	$4.82 \cdot 10^{-3}$
0.9500	0.9012	0.0012		$< 10^{-5}$	$2.39 \cdot 10^{-3}$

Chapter 11

Applications on Real Data

After reviewing the methods performance on simulated data, the recommended settings and parameters for the methods usage are demonstrated, so they provide reliable, unbiased and as accurate results as possible. This chapter describes the efforts that have been made on applying these methods on real-world data, assuming the real data represents a fractal set that can be modelled using the outlined techniques. The applications of the methods are demonstrated on following data:

- EEG data of patients with Alzheimer disease,
- econometric time-series of stock market indices development,
- biological autoluminiscence signal from mung beans,
- mammography screening images of patients with breast cancer.

The aim of the work is to prove that the fractal dimension or Hurst exponent could be a suitable feature for data classification and the chaotic information can be extracted using the newly proposed methods. Based on the input data, their length and quality we choose the most suitable method from the one described in previous chapters.

11.1 Application to EEG Signal

The rotational spectrum method has been proven to provide reliable estimation of correlation dimension for wide range of sets in Euclidean space, including the trajectories of time series. This method also allows us to study both the set in original time domain and in reconstructed space. The aim of our work is to analyse electroencephalogram (EEG) signals from patients that are suffering from Alzheimer disease (AD) and control normal (CN) patients. The subject of the analysis is 146 EEG samples of CN and 28 samples of AD patients. Every sample contains 19 time series of the signal development in time from 19 electrodes located on the patient head. The sample signal from 17th channel of CN patient is visualized in Fig.11.1, while the electrode scheme is visualised in Fig. 11.2. The EEG signal will be analysed in three different ways:

- as a vector process in 19-dimensional space,
- using state space reconstruction assuming it is a dynamic process,
- graph of each channel trajectory.

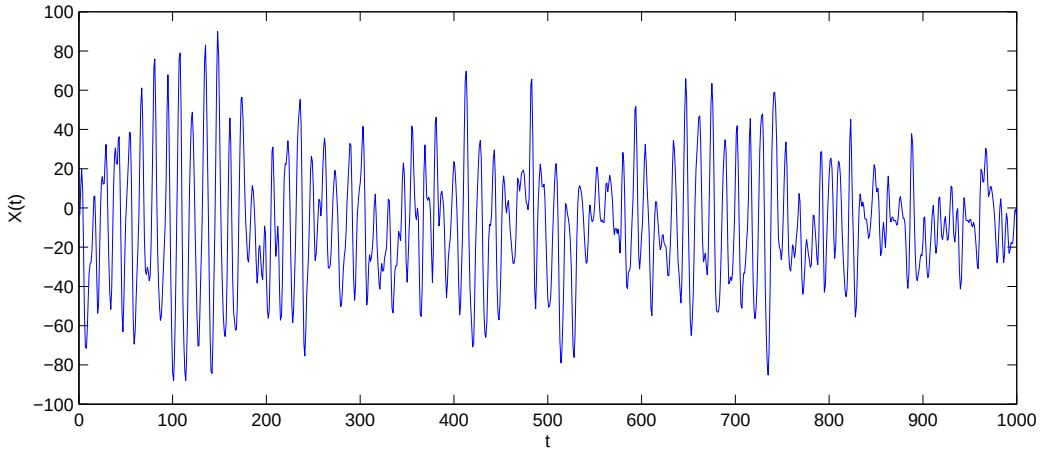


Figure 11.1: Sample of EEG signal from channel 17 of CN patient.

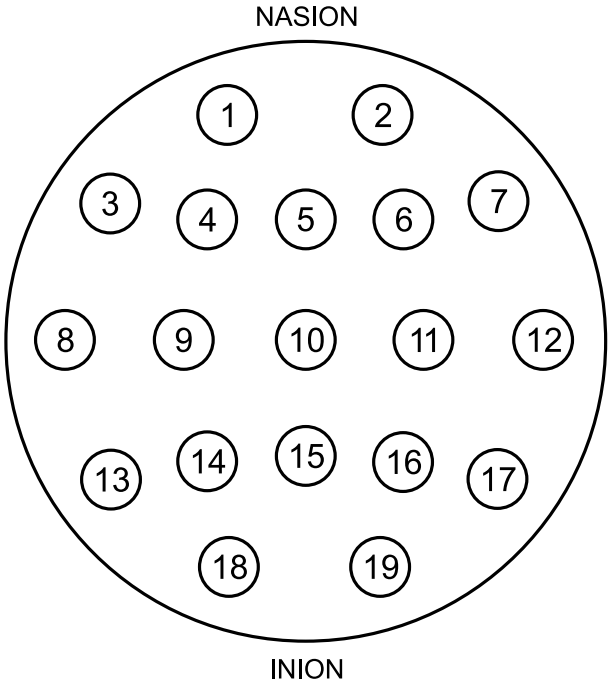


Figure 11.2: Electrode scheme 10-20.

Each of the time series coming from one of the 19 channels has the same length, $3 \cdot 10^5$ values. The aim of this study is to understand, whether there are channels that would have significantly different chaotic behaviour (by terms of measuring their correlation dimension) between the AD and CN patients.

For all calculations in this section, the kernel function H_∞ was used. The advantage of the rotational spectrum method is that it can analyse any finite dimensional set. Therefore, the first possibility of EEG signal analysis considers the signal as a vector process in 19-dimensional space and estimates the correlation dimension for each patient. This approach analyses the vector process in high dimensional space where for each time there is vector that represents the state of the whole human brain taking the information from all electrodes. We call this approach as global channel based analysis and this approach is denoted as G. The mean correlation dimension of the 19-dimensional process in case of CN patients was estimated as 4.3452 with standard deviation 0.9756 and in case of AD patients the dimension was 3.9058 with standard deviation 0.8458. The results from both groups were obeying normal distribution and by means of mean equality t-test the p-value was calculated as 0.0272. The results of the experiment are visualised in Fig. 11.3.

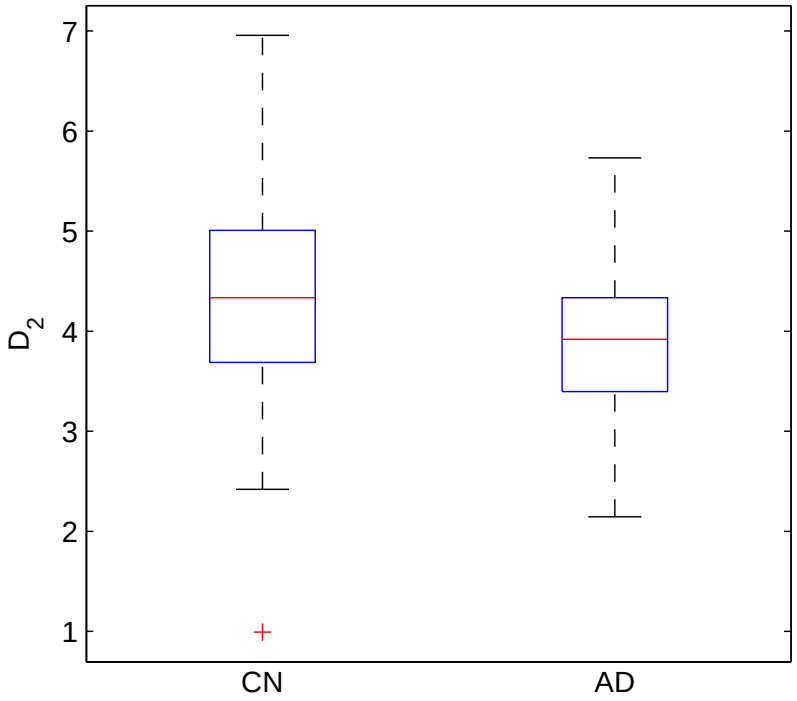


Figure 11.3: Global channel based analysis of EEG signal.

The second possibility of the investigation is to use state space reconstruction, considering the sample as discrete dynamical system. Considering the time series is a dynamic process, we can employ the Whitney embedding theorem for the estimation of attractor dimension. Recall that for reconstruction length $L \in \mathbb{N}$ and N-dimensional discrete dynamical system holds that if $L \geqslant 2N + 1$, then the correlation dimension D_2 of reconstructed attractor is the same as in the case of state series. Based on the analysis of EEG it was proven that the reconstruction length has to be greater than or equal 15. In case $L \geqslant 15$ the dimension estimates

vary only slightly and the results are not significantly higher as in the case of lower window lengths. Therefore, we use $L = 15$ for the analysis of the EEG channels. The state space reconstruction approach is denoted as SR.

The last approach for the EEG analysis is to estimate the correlation dimension of the graph of single trajectory. EEG time series graph is a set in two dimensional space, therefore the correlation dimension can vary between one and two. In this case, the input is the original time series measured by the electroencephalogram containing $3 \cdot 10^5$ points. This approach is denoted as graph method, GM.

Considering all three approaches, we performed 39 tests:

- 1 with G approach (global channel analysis),
- 19 with SR approach (space reconstruction),
- 19 with GM approach (graph method)

all with the null hypothesis

$$H_0 : D_{\mathrm{AD}} = D_{\mathrm{CN}} \tag{11.1}$$

where the D_{AD} is the expected value of correlation dimension of AD patients and D_{CN} is the expected value of CN patients. The aim is to find the channels where the difference between the dimension estimates is highest. Since in this case multiple testing is presented, we need to diminish the significance level based on Hochberg false discovery rate theorem to be $\alpha_{\mathrm{FDR}} = 0.00143$.

The estimates of correlation dimension were proven to have normal distribution along the groups, therefore we use the two-sample t-test for the hypothesis testing. Table 11.1 provides the results of the correlation dimension estimation together with the p-value for those channels, where the null hypothesis was rejected. The $\widehat{D}_{\mathrm{CN}}$ and $\widehat{D}_{\mathrm{AD}}$ are the mean values of correlation dimension in case of CN and AD patients, respectively, whereas std is their standard deviation. There was no significant result coming from the analysis of the graph method. The box-plots showing the differences between the two groups of patiens in case of global analysis is visualized in Fig. 11.3, whereas the results of the best option, state space reconstruction on 17th channel, is depicted in Fig. 11.4.

Table 11.1: Correlation dimension of EEG signal.

channel	method	$\widehat{D}_{\mathrm{CN}}$	$\mathrm{std}_{\mathrm{CN}}$	$\widehat{D}_{\mathrm{AD}}$	$\mathrm{std}_{\mathrm{AD}}$	p-value
-	G	1.7776	0.2760	1.6051	0.1282	0.0015
17	SR	5.5412	0.9385	4.8159	0.7146	0.0001
18	SR	5.5004	0.9484	4.8259	0.9733	0.0007
8	SR	5.4573	0.9181	4.8029	1.2555	0.0014
13	SR	5.4947	0.8995	4.9561	0.8167	0.0037
14	SR	5.6979	1.1058	5.1039	0.6296	0.0065

The most significant changes were recorded on the channels in the left occipital and the left temporal part of human brain where the null hypothesis about the equality

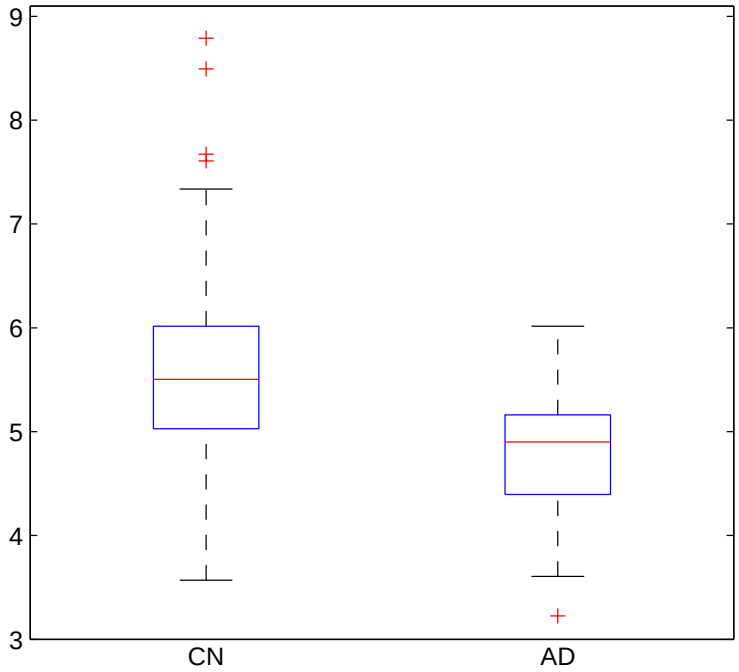


Figure 11.4: The best state space reconstruction (17th channel).

of mean values of correlation dimension was rejected. The channels with the lowest p-value are illustrated in Fig. 11.5.

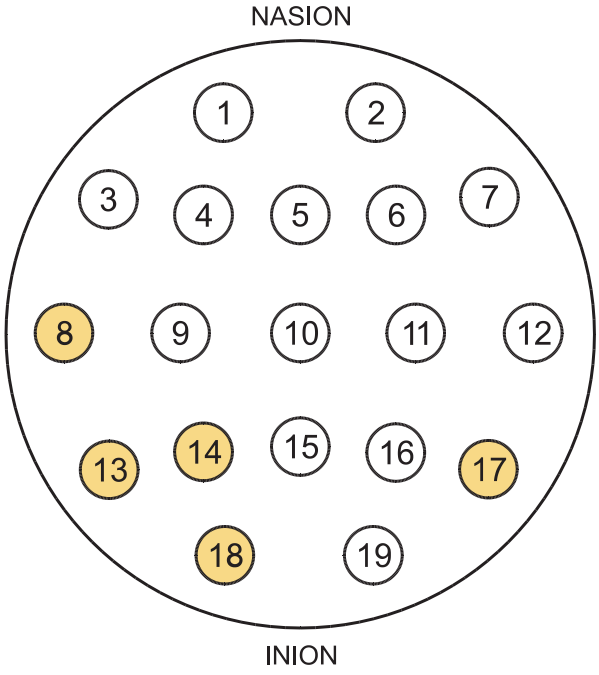


Figure 11.5: Significant channels from space reconstruction.

In comparison to graph method or global analysis, the more suitable technique for the EEG analysis is the state space reconstruction that was performed in our case in 15-dimensional space that brought significant differences between the two groups of patients. Statistically significant changes occurred on channels 17, 18 and 8. The experiments have proven that the correlation dimension is a powerful tool that can statistically distinguish between two groups of patients and provide small standard deviation of the estimates.

11.2 Application to Econometric Time Series

Any stock market index daily closing value can be considered as a realisation of a random variable using discrete time. The aim of this study is to examine the revisited zero-crossing method and analyse the autocorrelation and predictability of stock market indices. The study should reveal whether some stock market indices have different behaviour than others and describe how much their Hurst exponent varies in time. We used nine stock market indices for the analysis – CAC40, DAX, FTSE, HSI, NASDAQ, NIKKEI, SMI, SP500, TSX. The data contained the daily closing values of the stock markets between 2009-2016. For each index, there is roughly 260 records based on the number of trading days in that year. The aim is to find out which stock markets are more volatile in time and which ones are more predictable. We will use Hurst exponent as the measure of stock market variability.

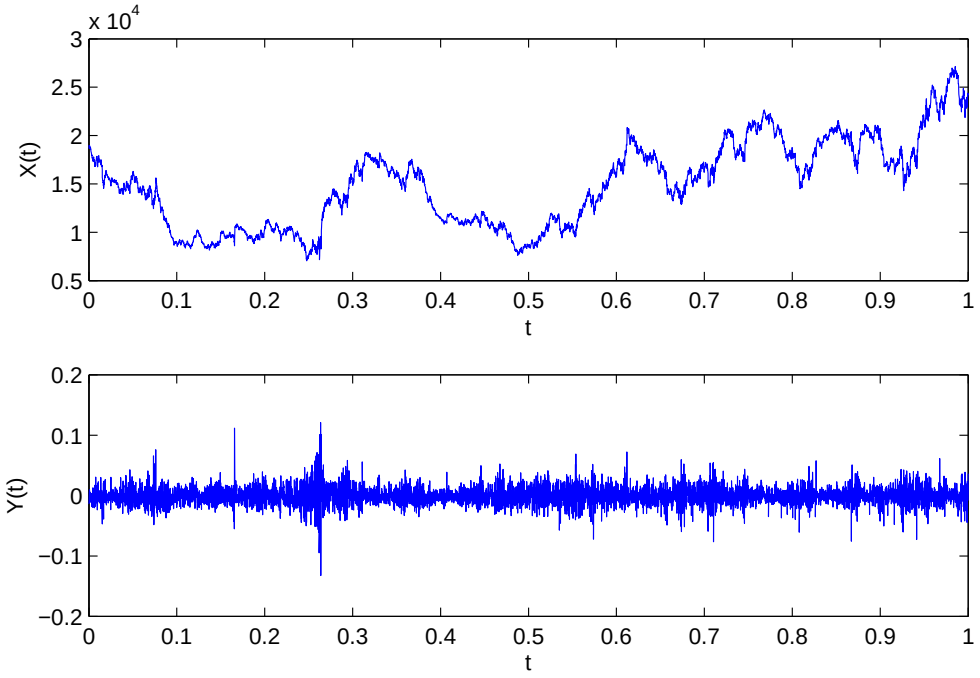


Figure 11.6: NIKKEI stock price development (top) and its logarithmic returns (bottom).

The theory of revisited zero-crossing and optimal segmentation is applied to the development of stock market indices. The data from stock market is first transformed

into their logarithmic returns as visualized in Fig. 11.6. We assume that the resulting time-series is a random process and model it using fGn properties of zero-crossing. At first it is necessary to investigate the behaviour of segmentation-dependent probability density. We have option to use natural Bayesian approach with $\alpha = 0$ or Jeffeys-Perks law with $\alpha = 1/2$. We divide the input time series into segments of different lengths L and using the posterior probability we can calculate the distribution $f_L(p)$of zero-crossing of the whole time series.

Probability density function $f_L(p)$ for TSX stock market index is depicted in Fig. 11.7 for $\alpha = 1/2$ and $L = 2$ (top left), 4 (top right), 7 (bottom left) and optimal segmentation $L^* = 12$ (bottom right). Recall, that the optimal segmentation is the lowest value of L when the $f_L(p)$ function is unimodal. Results for $\alpha = 0$ show similar behaviour, therefore we consider only the case $\alpha = 1/2$ in the following calculations.

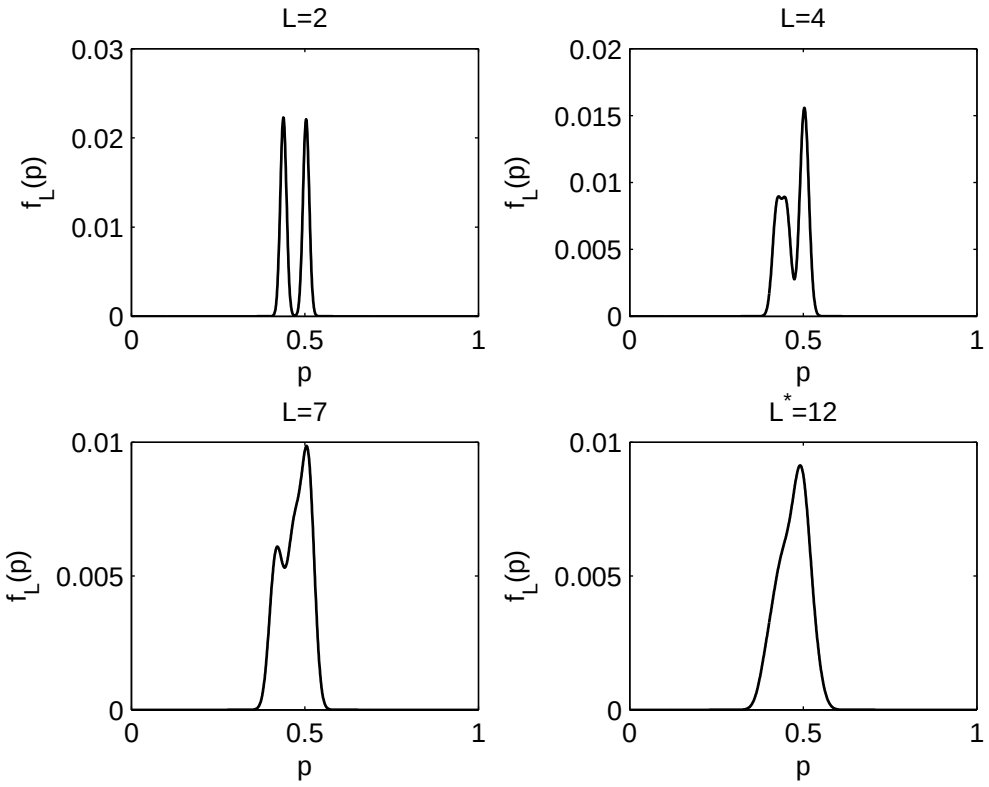


Figure 11.7: Bayesian density for various sizes of segmentation blocks of TSX.

With increasing number of segments L the support of function $f_L(p)$ extends while the total range decreases causing the unimodal property for large parameters L. It is generally not true that the function is unimodal for all parameters $L > L^*$. We demonstrate this fact in Fig. 11.8 showing the 95% confidence intervals and mean values of H for all feasible divisions with unimodal $f_L(p)$. The optimal segmentation occurs for $L^* = 5$, however, the function has not unique peak for all $5 < L < 10$. Once the segmentation on L intervals is fine enough, the new additions to segmentation-dependent probability distribution are no longer significant to spoil

the trend of the whole function obtained by dividing into $L-1$ blocks.

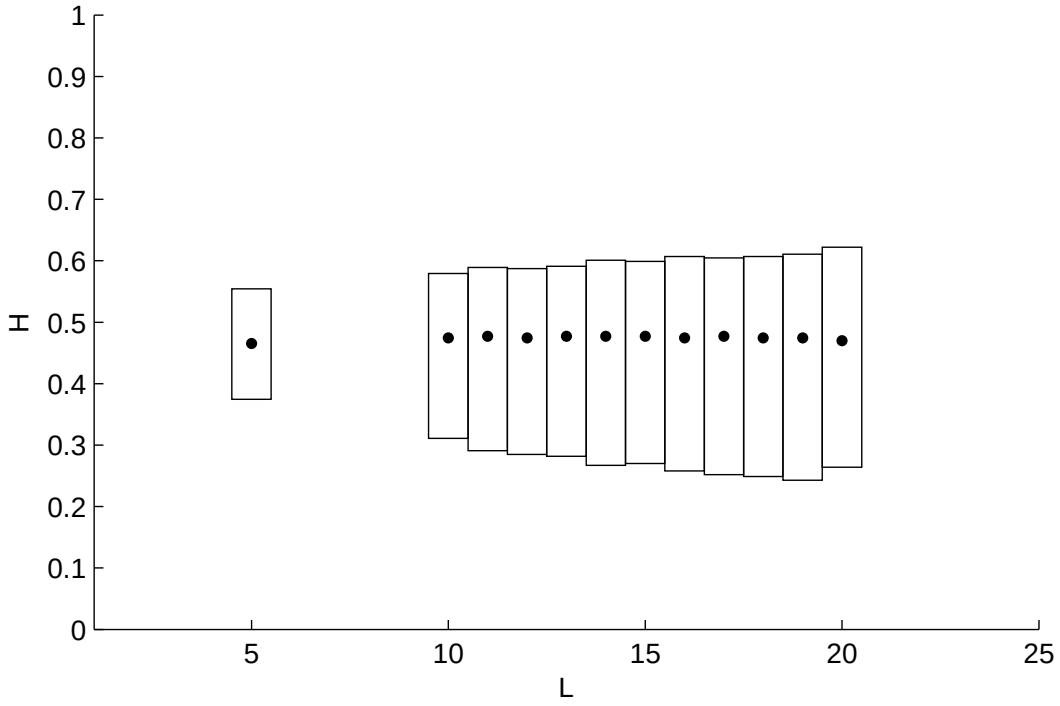


Figure 11.8: 95% confidence intervals for feasible segmentations of CAC40.

Table 11.2 shows the traditional (ZC) point estimate (7.11) of H together with expected values and 95% confidence intervals obtained by Bayesian approach without segmentation (for $L = 1$) and revisited approach showing the parameter of optimal segmentation L^* and appropriate statistical characteristics. When the parameter of optimal segmentation equals 1 (for DAX,HSI and SMI), it means the function $f_L(p)$ is unimodal from the beginning signifying more or less constant parameter H in the entirety of the signal. NASDAQ and TSX are the stock markets whose Hurst parameters varies in time a lot, because their optimal segmentation parameter equals 14 and 12, respectively.

The revisited zero crossing provides usually wider confidence intervals for the majority of stock markets than the traditional approach. However, the optimal expected values remain almost the same in comparison with the case without segmentation. In most cases, larger confidence interval indicate that the specific index does not follow the same fractal behaviour during the entire investigated period. Therefore the method is able to find new fractal patterns by means of dividing original signal – when the optimal segmentation is reached, the estimates are more realistic and therefore useful for further analysis such as investment recommendations or predictions.

Table 11.2: Comparison of traditional and revisited method.

index	ZC	no segmentation ($L = 1$)		optimal segmentation		
	H^*	EH	95% CI	L^*	EH	95% CI
CAC40	0.4745	0.4672	(0.4435; 0.4909)	5	0.4652	(0.3746;0.5545)
DAX	0.5007	0.4790	(0.4556; 0.5023)	1	0.4790	(0.4556;0.5023)
FTSE	0.5502	0.4906	(0.4676; 0.5135)	3	0.4863	(0.4361;0.5482)
HSI	0.5664	0.4974	(0.4746; 0.5201)	1	0.4974	(0.4746;0.5201)
NASDAQ	0.5631	0.5584	(0.5375; 0.5792)	14	0.5418	(0.4008;0.7136)
NIKKEI	0.4583	0.4553	(0.4312; 0.4793)	3	0.4532	(0.4034;0.5068)
SMI	0.5303	0.5087	(0.4863; 0.5310)	1	0.5087	(0.4863;0.5310)
SP500	0.4722	0.4456	(0.4212; 0.4699)	6	0.4460	(0.3225;0.5628)
TSX	0.6107	0.5646	(0.5439; 0.5852)	12	0.5607	(0.3957;0.6992)

11.3 Application to Autoluminiscence Signal

This section describes analysis of signal from biological autoluminiscence (BAL). Autoluminiscence is a weak light that is generated practically by all organisms and is present in the 300-700 nm wavelength range, therefore the light cannot be spotted by naked human eye. The emitting mechanism is not known yet, but widely accepted hypothesis is that BAL is related to a chamical generation of electron-excited states of biomolecules in the course of oxidative metabolism and oxidative stress. This kind of signal arises from metabolic and physiological processes and can be detected using photomultiplier in the form of emitted photons. The object of our study is the BAL signal of mung beans and the aim is to find out whether there is autocorrelation or dependency in the signal.

The biological material in this case were mung bean seeds that were firstly sterilized, disinfected and having the green covers removed. The photon emission was measured in low-noise photomultiplier tube (PMT). A PMT can exhibit noise on its own, therefore it was necessary to introduce a methodology, that could distinguish between the noise signal of PMT and the signal from the mung beans. For this work we generated the reference signal as a sum of measured detector noise and computer-generated Poisson signal with given λ in every experimental point where $\lambda = \mathrm{E}\, y_k^{\mathrm{B}} - \mathrm{E}\, y_k^{\mathrm{D}}$, where y_k^{B} and y_k^{D} are signal mean values of mung beans and noise, respectively. The respective values of λ in case of 200μs signal as well as 500 μs signal are calculated in Tab. 11.3. Hence, experimentally detected BAL signals from mung beans and reference signals have practically the same mean value and same signal-to-noise ratio.

Since the input series are rather short, we will be using the method for short time series to investigate the statistical properties of BAL. To sum up, for the analysis, we have three types of signals available:

- (B) - mung beans signal y_k^{B} (containing PMT noise),
- (D) - noise signal of PMT detector y_k^{D},

- (R) - reference signal as a sum of measured detector noise (D) and computer-generated Poisson noise denoted as $y_k^{\mathbb{R}}$

Table 11.3: Mean values of mung beans signal and noise.

T_s	200 μs	500 μs
E y_k^{B}	0.0115	0.0288
E y_k^{D}	0.0036	0.0088
λ	0.0079	0.0200

The physical data collection happens in a photomultiplier tube which has two bin size settings used to collect the signals: $T_{\mathrm{s}} = 200\mu$s and 500μs. For each sampling period, we have corresponding mung bean signals, detector noise signals, and computer-generated reference signals. Both mung beans signal and PMT detector noise signal are assumed to be stationary with their mean values with the Poisson distribution. Therefore, they can be represented by their mean values E y_k^{B} and E y_k^{D} that are estimated from the measured data.

Recall the aim of study is to compare mung beans signal with the reference signal and find statistical difference between them using their autocorrelation. With each of these two signals independently, we performed basic data processing. This procedure describes the normalization of the data, which is the essential property of fBm processes. At first the input time series y_k for $k = 0, 1, \ldots, Q-1$ was cumulatively summed for a window size $h \in \mathbb{N}$ and Anscombe transformation [84] was performed. The resulting signal z_k can be expressed based on the output from measuring device y_k as

$$z_k = 2 \cdot \left(\frac{3}{8} + \sum_{i=kh}^{(k+1)h-1} y_i \right)^{1/2} \tag{11.2}$$

for $k = 0, \ldots, M-1$. This transformations assures stationarity by terms of variance and guarantees Gaussian distribution of the resulting signal. The example of the input signal y_k from mung beans with $T_{\mathrm{s}} = 500\mu s$ and corresponding z_k is visualized in Fig. 11.9.

Having signal from the mung beans photon emission as well as the reference signal, we will use likelihood ratio test [85] to decide, whether the Hurst exponent of both samples is significantly different. We denote H_{D} as the Hurst exponent estimate of the PMT detector noise or reference signal and H_{B} as the Hurst exponent estimate of mung emission. The overall error (sum of the squares of residuals) is defined as

$$SSQ_{\mathrm{FULL}} = \sum_{i=1}^{M} \sum_{j=1}^{N-1} \left(\rho_{i,j}^{\mathrm{B}} - \rho_j(H_{\mathrm{B}}) \right)^2 + \sum_{i=1}^{M} \sum_{j=1}^{N-1} \left(\rho_{i,j}^{\mathrm{D}} - \rho_j(H_{\mathrm{D}}) \right)^2, \tag{11.3}$$

where $\rho^{\mathrm{D}}, \rho^{\mathrm{B}}$ are the autocorrelation coefficient of the noise and photon emission, respectively. The case of $j = 0$ is excluded due to $\rho_{i,0}^{D} = \rho_0(H_{\mathrm{D}}) = 1$ for all

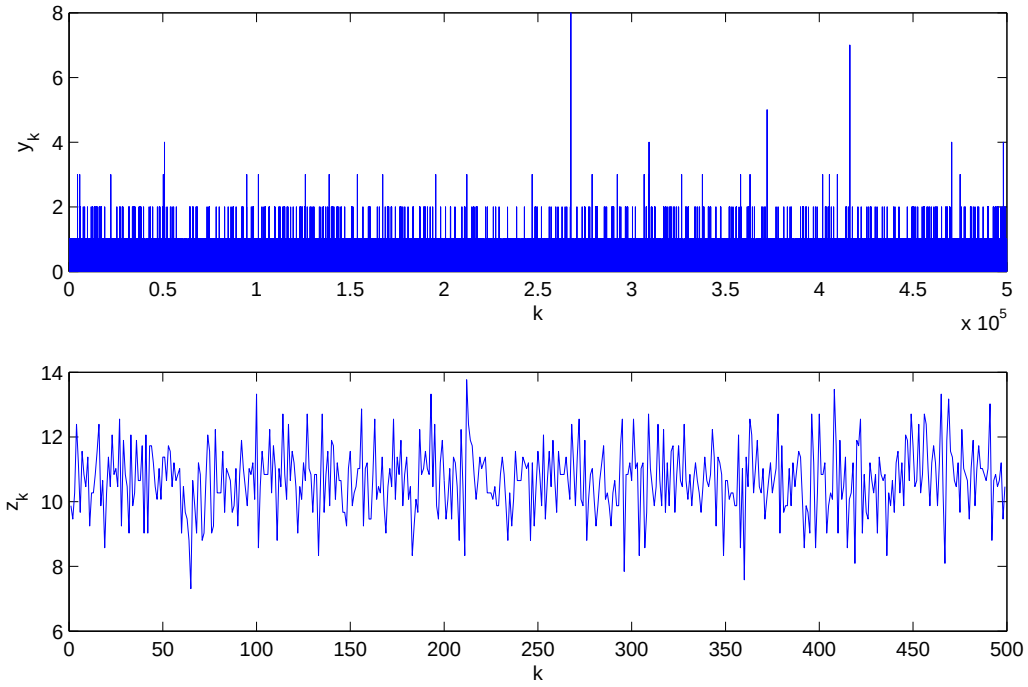


Figure 11.9: Mung beans signal y_k (top) and corresponding z_k (bottom).

$i = 1, .., M$. Using sub-model satisfying $H_{\mathrm{B}} = H_{\mathrm{D}}$ we get

$$SSQ_{\mathrm{SUB}} = \sum_{i=1}^{M} \sum_{j=1}^{N-1} \left(\rho_{i,j}^{\mathrm{B}} - \rho_j(H_{\mathrm{D}})\right)^2 + \sum_{i=1}^{M} \sum_{j=1}^{N-1} \left(\rho_{i,j}^{\mathrm{D}} - \rho_j(H_{\mathrm{D}})\right)^2. \tag{11.4}$$

Using likelihood ratio (LR) test of significant difference between the sub-model and the full model, we calculate

$$\chi^2 = 2 \ln \frac{L_{\mathrm{FULL}}}{L_{\mathrm{SUB}}} = M \cdot (N-1) \cdot \ln \frac{SSQ_{\mathrm{SUB}}}{SSQ_{\mathrm{FULL}}}, \tag{11.5}$$

where L_{FULL} and L_{SUB} are corresponding likelihoods.

When the hypothesis $\mathrm{H}_0 : H_{\mathrm{D}} = H_{\mathrm{B}}$ holds, i.e. the full model has the same validity as the submodel, the criterion has χ^2_1 distribution due to single parameter constrain.

There is no prior knowledge of optimal model length, accumulation compression, and Hurst exponent. Therefore, we will apply the maximum likelihood method of Hurst exponent estimation for the various model and segment lengths, and then we will individually test the differences in the Hurst exponent. However, there is a finite number of reasonable pairs (model length N, segment length h), which will cause the phenomenon of the multiple hypothesis testing. After the False Discovery Rate (FDR) correction, we will localize the model and segment lengths, which cause significant differences in the Hurst exponent. These pairs (h, N) will be declared as significantly sensitive to the signal differences in the Hurst exponent.

Having signals with two different bin sizes, we will use the signal bin size $T_b = 200\mu$s as a training set and the signal with $T_b = 500\mu$s as a verification set. Normalized

mung beans and reference signals with bin size $T_b = 200$ μs and length $Q = 500\,000$ were the subject of the initial analysis. The signal accumulation of size h was applied to the signals, therefore the number of bins was $\lfloor Q/h \rfloor$. After the accumulation, the signal is divided into segments of length N. Due to the memory of fBm process, we will use only the odd segments for the calculation of autocorrelation function and the even segments are excluded. The new signal has length $\lceil \lfloor \lfloor Q/h \rfloor / N \rfloor /2 \rceil$. Using maximum likelihood method, we obtain the corresponding H_{D} and H_{B} estimates for the Hurst exponent of referential signal and mung beans, respectively. Based on these estimates, we can derive the p-values of LR test using (11.5) statistics.

In our case, we performed altogether $11 \times 11 = 121$ tests for $h = 1500, 1550, \ldots, 2000$ and $N = 20, 21, \ldots, 30$. Accumulation h could not be higher than 2000 due to the rapid decrease of the number of processed segments. The values $h < 1500$ caused lower event frequencies and the conversion from Poisson noise to Gaussian noise is not guaranteed. Similar reasons are for the range of parameter N. In fact, the fractional model is less discriminative for $N < 20$ and the case $N > 30$ reduces the number of segments. Due to multiple testing and obeying the Hochberg-Benjamini principle, we diminish the significance level from 0.05 to $\alpha_{\mathrm{FDR}} = 0.000050$. The p-values as decadic logarithms are shown in Tab. 11.4.

Table 11.4: Difference between the estimated Hurst exponent of mung beans (B) and reference signal (R) as $(-\log_{10} p)$-values of likelihood ratio test (11.5).

$h \setminus N$	20	21	22	23	24	25	26	27	28	29	30
1500	1.188	2.934	1.835	1.888	3.175	1.645	2.284	0.863	0.506	0.192	0.762
1550	1.172	1.617	0.394	0.887	1.420	1.470	0.912	2.113	1.651	0.026	0.691
1600	1.978	1.576	0.646	0.523	1.217	0.394	1.597	0.786	1.487	0.880	1.859
1650	0.990	1.616	1.127	2.024	1.209	0.651	1.635	0.909	1.906	3.573	2.927
1700	0.772	0.621	1.288	1.196	1.239	0.488	0.407	1.175	2.658	0.463	0.776
1750	1.475	2.325	1.269	3.131	**4.638**	1.535	2.370	1.017	0.726	0.412	1.945
1800	0.465	1.455	1.394	1.098	1.313	0.180	2.661	2.064	2.449	1.917	2.001
1850	2.377	2.010	1.308	0.567	1.533	2.382	3.184	**4.301**	3.328	2.418	1.968
1900	2.599	0.879	0.850	0.629	1.053	1.264	0.950	0.943	1.397	2.093	0.142
1950	2.574	0.095	0.706	1.900	2.843	2.874	3.261	2.514	3.462	2.501	2.405
2000	2.212	1.611	1.315	0.935	1.040	1.232	0.922	0.282	0.366	1.159	0.963

In these settings, there were two cases where the Hurst exponent was significantly different. The results from these two cases are displayed in Table 11.5. The 95% interval of Hurst exponent estimates of mungo beans signal (H_{B}) was [0.2108, 0.5086], while the 95% interval for reference signal (H_{R}) was [0.4041,0.5931].

The lowest p-value was obtained in the case of $(h, N) = (1750, 24)$, which represents the segmentation into bins with duration 1750×200 μs $= 350\,000$ μs $= 0.35$ sec.

As the verification set, the signal with $T_{\mathrm{s}} = 500$ μs was taken into account, following the same procedure as the previous one. The accumulation parameter h was accordingly diminished to 2/5 of its previous value to guarantee the same segment length.

We perform the verification for the combination of signals (B) and (R) similarly as

Table 11.5: Estimated Hurst exponent values for mung beans (B) signal and reference signal (R).

h	N	H_{B}	H_{R}	p-val	$-\log_{10} p$-val
1750	24	0.4142	0.5299	2.3135×10^{-5}	4.638
1850	27	0.3569	0.4291	4.9977×10^{-5}	4.301

in the previous case and additionally for the combination of (B) and (D). The first set of signals ((B) and (R)) will be used to test if the photon emission is not random and has a negative memory, while the results from the second set ((B) and (D)) of signals will be used to test if there is a significant difference between the cases, when the PMT detects BAL signals from mung beans compared to PMT noise. We use the significant cases from Table 11.5 to estimate their Hurst exponent and the results on verification set is displayed in Tab.11.6. The parameter value h=700 for 500 μs signals corresponds to h=1750 for 200 μs signals. The variables s_1, s_2 denote the pair of signals, whereas the H_{X} denotes the estimation of Hurst exponent of the signal s_2.

Table 11.6: Estimated Hurst exponent values from verification dataset.

s_1	s_2	h	N	H_{B}	H_{X}	p-val
B	R	700	24	0.4032	0.4415	0.0130
B	R	740	27	0.3761	0.4112	0.0042
B	D	700	24	0.4032	0.4378	0.0169
B	D	740	27	0.3761	0.4480	0.0054

We performed four tests, and according to Hochberg-Benjamini false discovery rate, we diminish the $\alpha_{\mathrm{FDR}} = 0.0169$. Therefore, all four cases are considered significant, and we reject the hypothesis that the Hurst exponent of mung beans would be the same as H_{X}.

For comparison, we also performed a similar analysis with noise signal (D) and mungo beans signal (B) and captured the results in the Table 11.7. Using Hochberg-Benjamini principle, there is only one combination (h,N) = (1850,27) that is significant.

Results from statistical analysis and testing suggest that the mung beans signal has a negative memory and its Hurst exponent is lower than the referential signal. The 0.35s ($700 \cdot 500\mu s$) was the time with statistically significant differences between the mung bean signal and reference signal. This could correspond to the rate of underlying chemical process that give rise to BAL. The methodology of signal preprocessing via cumulative sum and Anscombe transformation has assured that the signal can be modelled using the fractional processes. Therefore it is recommended as a useful technique on how to process input data for subsequent analysis using fBm and fGn models.

Table 11.7: Difference between the estimated Hurst exponent of mung beans (B) and noise signal (D) as $(-\log_{10} p)$-values of likelihood ratio test (11.5).

$h \setminus N$	20	21	22	23	24	25	26	27	28	29	30
1500	2.181	2.881	1.395	1.656	3.192	2.058	2.203	0.087	0.444	1.131	3.319
1550	2.191	1.125	0.446	0.736	0.348	2.101	0.283	2.378	1.936	1.560	0.407
1600	2.008	2.039	1.259	0.006	1.071	1.098	1.604	2.003	0.665	1.468	1.415
1650	1.230	1.535	1.334	2.804	1.249	0.914	0.858	1.470	2.923	2.190	3.056
1700	1.015	1.359	0.376	0.460	0.473	0.665	0.182	0.906	1.812	1.532	0.107
1750	0.769	1.218	0.926	0.346	1.771	2.386	0.687	0.331	1.286	1.136	1.605
1800	0.696	1.791	0.871	0.666	2.058	2.153	2.053	1.145	2.400	0.692	0.881
1850	2.891	1.967	1.349	0.297	1.051	2.343	2.754	**4.602**	3.472	0.779	1.432
1900	2.795	0.434	0.708	1.273	0.619	0.631	1.797	1.101	0.395	1.371	1.069
1950	1.657	1.694	0.386	2.885	2.373	2.199	2.504	2.292	1.651	0.926	2.209
2000	1.281	0.048	0.215	1.050	0.140	1.031	0.234	1.303	1.186	0.239	0.939

11.4 Application to Mammography Images

As an application of fractal theory on images, we want to study mammography screening images, assuming the scan of woman's breast is a sample of fractional Brownian surface. The introduced methodology of Wishart estimate of Hurst exponent can be tailored to analyze surfaces in the form of intensity images. The data are taken from the Digital Database for Screening Mammography [86, 87] containing public and open-access data for research in image processing. For the study, we have selected 140 patients who have benign lump (BL) and 140 patients with cancerous lump (CL). The cranio caudal (CC) images are used for the purposes of this study. The images were captured by DBA M2100 ImageClear digitizer with sampling rate 42 microns and 16 bits gray level. The pictures were processed in Matlab as grayscale pictures and cropped, so the picture only contains human tissue without background noise.

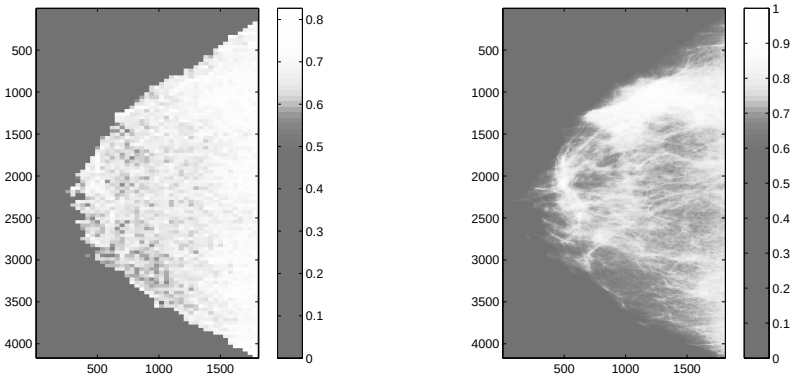


Figure 11.10: Hurst exponent distribution map (left) and the original intensity image (right) of a CC image of woman's right breast.

Afterwards, the threshold technique has been applied to the images. For image with

dimensions $P, Q \in \mathbb{N}$, we define dimensionless parameter $\theta \in (0; 1)$ and define our region of interest as

$$ROI = \{(i, j) : x_{i,j} > \theta \cdot \max_{\substack{1 \leqslant i \leqslant Q \\ 1 \leqslant j \leqslant P}} x_{i,j}\} \tag{11.6}$$

to localize the benign and cancerous lumps on the breast image. For our experiment, we have chosen three thresholds selecting points with high intensity from the original x-ray scans. The threshold $\theta = 0.80$ corresponds to roughly 5%, threshold $\theta = 0.85$ assures roughly 3% and threshold $\theta = 0.9$ corresponds to roughly 1% of pixels of the original image. For generating scatter matrix, we will use symmetric masks with radius from three to seven pixels, i.e. $\rho \in \{3, 4, 5, 6, 7\}$.

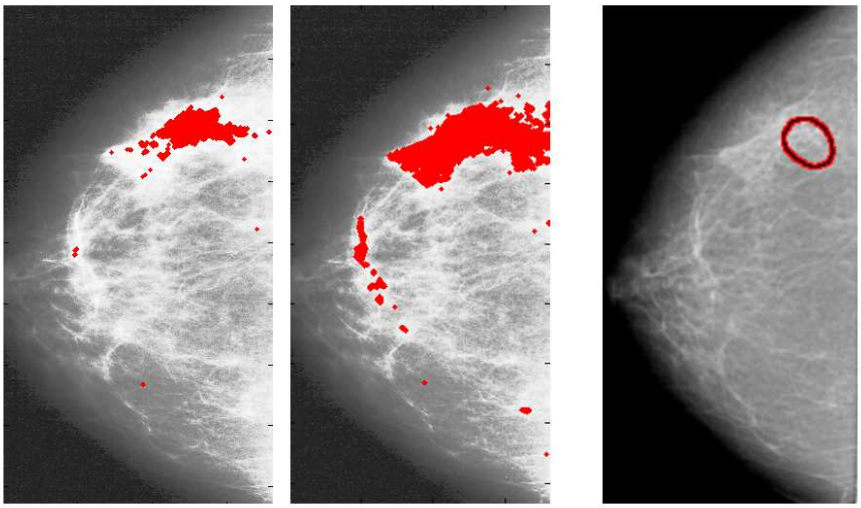


Figure 11.11: Regions selected for analysis for benign finding for $\theta = 0.9$ (left), $\theta = 0.85$ (middle) and affected area highlighted by a medical doctor (right).

We will use two approaches to compare the results between BL and CL patients. First approach estimates H for each patient in the respective groups. Using Wilcoxon ranksum test, we will test hypothesis

$$H_0 : H_{\mathrm{B}} = H_{\mathrm{C}}, \tag{11.7}$$

where H_{B} and H_{C} are the Hurst exponent estimates of BL and CL groups, respectively. The results of the estimation are displayed in Tab. 11.8. Using false discovery rate, we diminish the critical level of significance to $\alpha_{\mathrm{FDR}} = 0.0217$. The cases, where the null hypothesis is rejected are highlighted in bold.

The case with highest significance (lowest p-value) in the traditional approach is $(\theta, \rho) = (0.85, 5)$. This combination will be investigated further using another approach that employs the estimation of both parameters (H, σ) and the hypothesis claims, that the observed images of both cancerous and benign patterns are samples of fractional Brownian surface, but they can vary in parameters H and/or σ. We denote H_B as the Hurst exponent of the BL patients and H_C as the Hurst exponent of the CL patients and the corresponding σ as σ_C and σ_B respectively. Therefore, there are four sub-models in run:

Table 11.8: Wilcoxon-Mann-Whitney test for various thresholds and masks.

θ	ρ	H_{B}	std(H_{B})	H_{C}	std(H_{C})	p-val
0.8	3	0.5513	0.1912	0.5663	0.1891	0.1560
0.8	4	0.5943	0.1768	0.6281	0.1680	0.0732
0.8	5	0.5833	0.1766	0.6172	0.1733	0.0623
0.8	6	0.6384	0.1885	0.6664	0.1744	0.1386
0.8	7	0.6225	0.1866	0.6518	0.1737	0.1341
0.85	3	0.5313	0.1972	0.5782	0.2258	**0.0142**
0.85	4	0.6195	0.1752	0.6645	0.1927	**0.0107**
0.85	5	0.6158	0.1803	0.6601	0.1956	**0.0104**
0.85	6	0.6793	0.1917	0.7179	0.2029	**0.0192**
0.85	7	0.6616	0.1887	0.7018	0.2025	**0.0217**
0.9	3	0.4964	0.2454	0.5102	0.2781	0.0634
0.9	4	0.6551	0.1877	0.6908	0.2106	0.0245
0.9	5	0.6586	0.1895	0.6961	0.2199	**0.0212**
0.9	6	0.7300	0.2078	0.7549	0.2190	0.0445
0.9	7	0.7084	0.2053	0.7378	0.2178	0.0476

1. model $H = H_B = H_C$ and $\sigma = \sigma_C = \sigma_B$
2. model $H = H_B = H_C$ and σ_C, σ_B
3. model H_B, H_C and $\sigma = \sigma_C = \sigma_B$
4. model H_B, H_C and σ_C, σ_B

Their hierarchy is depicted on Fig. 11.12. Both H and σ parameters can vary due to structural changes of the tissue. The estimated values of parameter σ also depend on X-ray intensity and sample depth, therefore their differences are of low practical importance. The estimated values of parameter H also depend on scanning apparatus resolution, which is constant for given data set and therefore more suitable for biomedical interpretation. The results of parameters for each of the models separately is presented in Tab. 11.9.

The models will be compared using likelihood ratio test. We can perform 4 comparisons: 2 to 1 (comparison D), 3 to 1 (comparison C), 4 to 2 (comparison B) and 4 to 3 (comparison A), while the first mentioned model in the pair is called full model and the latter one is its submodel. Having four comparisons in place allows us to measure, whether the difference between the BL and CL is in the Hurst exponent H or in the σ. For each of the models, we estimate its parameters by maximizing (9.48) and then calculate corresponding logarithm of likelihood as

$$\ln L = \ln f(\mathbb{S}, \mathbb{V}, p, m) \tag{11.8}$$

using (9.45). Comparing any model to its submodel, we calculate the log-likelihood of the full model and denote it $\ln L_{\mathrm{FULL}}$, while the log-likelihood of the submodel

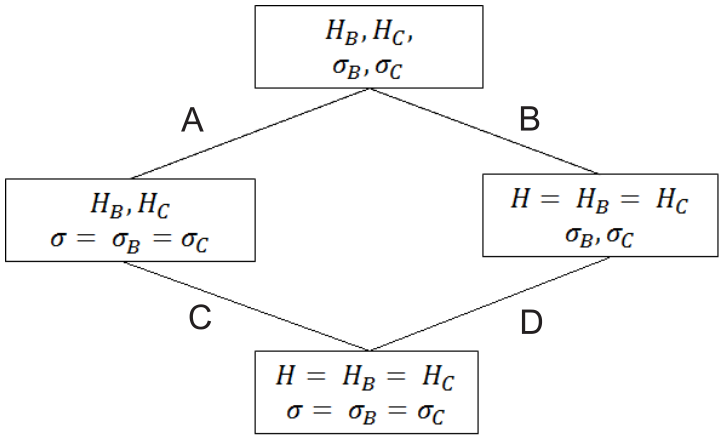


Figure 11.12: Model hierarchy schema.

equals $\ln L_{\mathrm{SUB}}$. In our setting, any full model always differs in one free parameter in comparison to its sub-model. It holds that

$$2\left(\ln L_{\mathrm{FULL}} - \ln L_{\mathrm{SUB}}\right) \sim \chi^2. \tag{11.9}$$

Table 11.9: Estimates of model parameters.

model	H_{B}	H_{C}	σ_{B}	σ_{C}
1	0.5037	0.5037	0.0073	0.0073
2	0.5034	0.5043	0.0070	0.0075
3	0.5072	0.5085	0.0073	0.0073
4	0.4980	0.5085	0.0070	0.0075

Using the likelihood ratio tests, the comparisons A,B,C,D were evaluated, the χ^2 value together with the logarithms of p-values were calculated. The comparisons are depicted in tab 11.10. All four sub models significantly differ in the sense of likelihood ratio test and therefore the fourth model i.e. the full model statistically dominates the others. It implies the rejection of hypothesis $H_{\mathrm{B}} = H_{\mathrm{C}}$.

Table 11.10: Model and sub-model comparison.

full-model	sub-model	comparison	χ^2	$\ln p$-val
4	3	A	$1.13 \cdot 10^5$	$-5.63 \cdot 10^4$
4	2	B	$1.68 \cdot 10^3$	$-8.46 \cdot 10^2$
3	1	C	$1.53 \cdot 10^5$	$-7.63 \cdot 10^4$
2	1	D	$4.16 \cdot 10^4$	$-2.08 \cdot 10^4$

Both Wilcoxon-approach combined with the false discovery rate as well as the likelihood ratio test have discovered sigificant differences between the CL and BL patients. The image of CL patients has statistically higher Hurst exponent when

considering the tissue has properties of fractional Brownian surface. The distribution of Hurst exponent in case of CL and BL for $(\theta, \rho) = (0.85, 5)$ is visualized using boxplot in Fig. 11.13.

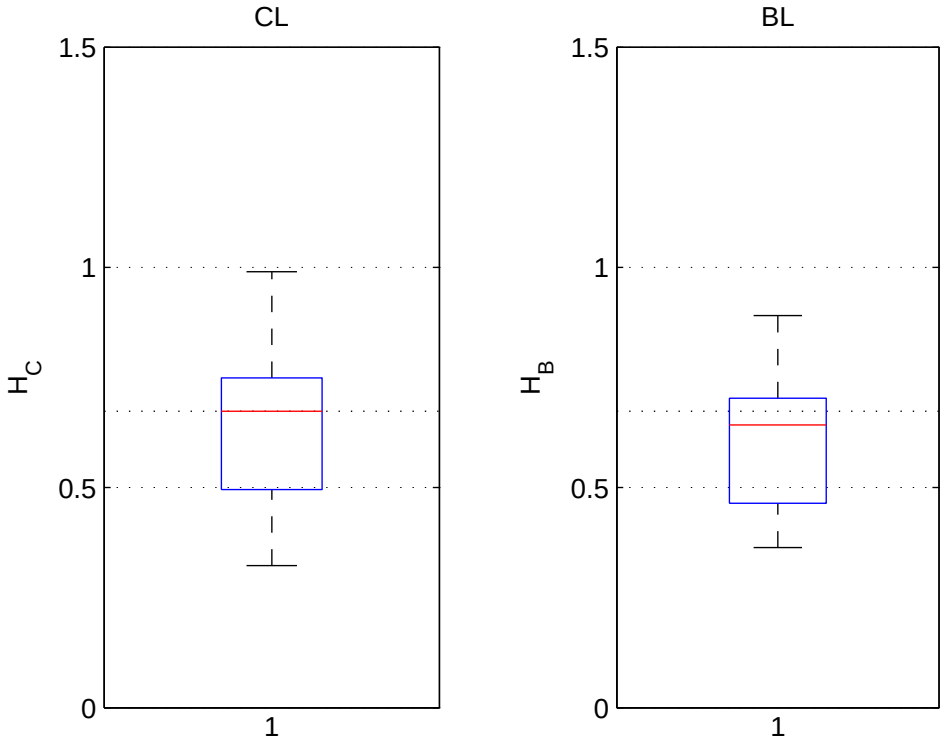


Figure 11.13: Distribution of Hurst exponent in case of BL (right) and CL (left) patients.

The methodology of Hurst exponent estimate was applied to cranio caudal mammography images, considering the human tissue as a sample of fractional Brownian surface with unknown Hurst exponent. Two approaches involved testing through false discovery rate combined with the Wilcoxon test as well as likelihood ratio test. Both approaches have rejected the null hypothesis that the CL and BL patients have the same Hurst exponent, resulting in patients with cancerous findings having higher Hurst exponent and more chaotic structures in the investigated breast. This is in accordance with the latest research findings in mammogram diagnostics [88].

Conclusions

There were five main goals to be achieved in this thesis;

- development of a new method for correlation dimension estimation,
- development of a new method for Renyi dimension estimation,
- development of a new method for Hurst exponent estimation,
- application of the methods to artificial sets with known dimension,
- application of the methods to biomedical and econometric signals.

The first introduced novel method of rotational spectrum can be used as an alternative technique for correlation dimension estimation. The method utilizes the rotation of a power spectrum of a set around the origin and can be expressed in an explicit formula using the expected value operator and a kernel function. The rotational spectrum has smoother dependency than standard correlation sum and can provide more accurate estimate of correlation dimension. Moreover, the ratio of logarithm of rotational spectrum and logarithm of the frequency can be used for correlation dimension estimation in similar way as in the case of correlation sum approach. The method has been published in journal Chaos, Solitons & Fractals.

The estimation using modified Renyi entropy is alternative method to Renyi dimension estimation. The method utilizes the Parzen density estimation based on the points from the sample set and defines a new entropy that is translational and rotational invariant. The resulting entropy can be used for an alternative definition of the Renyi dimension for any parameter $\alpha \geqslant 0$. Additionally, it was proven that the modified Renyi dimension coincides with the definition of both capacity and correlation dimension for any measurable point set. Via simulations, the validity of the definition was proven also for other dimension types. The method has been published in journal Chaos, Solitons & Fractals.

By terms of Hurst exponent estimation, the focus was both on estimation from short time series as well as from longer signals and fractional surfaces. The differenced fractional Brownian bridge is a generalization of Brownian bridge that is especially designed to estimate Hurst exponent from short time series. The autocorrelation function of the resulting signal was theoretically derived and is used for the parameter estimation. The method has been published in journal Signal, Image and Videoprocessing. The autocorrelation properties were used for derivation of another

method for Hurst exponent estimation using Wishart distribution. The method utilizes the input fBm division into smaller samples and calculates the aggregate scatter matrix that has known probability density function. Using maximum likelihood approach and based on the fact that the autocorrelation of the original process is known, one can estimate the Hurst exponent of the whole sample. The method can be used both in one-dimension as well as for fractional surfaces and masses. The method has been published in journal Physica A.

The methods of estimation have been successfully verified in the experimental part. We have used 10 parametrizable sets with known Renyi dimension to validate the rotational spectrum, modified Renyi entropy as well as Hurst exponent estimation methods. The method's performance has been examinated both in comparison to traditional approaches as well as artificial data via statistical testing. To the main outcomes of the statistical testing belongs:

- **comparison between the novel method and standard method**; The methods were compared against traditional approaches of dimension estimation (correlation sum, box-counting, ...) to verify the performance. It was found out that the proposed methods are not worse than the standard approach in the whole range of the eligible parameters; moreover, the proposed metrics has smoother dependency than the traditional ones and frequently result in a more accurate estimate with smaller standard deviation of the estimate.

- **recommended parameter setting**; The newly introduced methods are designed with a set of parameters (number of samples, size of input data, frequency range, ...). Using large number of examples, it was possible to recommend the best parameter setting where the methods provide unbiased estimation of the dimension. The parameter recommendation is important, since it can be later used for investigation of signals with unknown dimension or unknown Hurst exponent.

- **unbiasedness of the estimation**; Thanks to the fact that the fractal dimension of simulated sets is known, the statistical testing has revealed in which range of dimensions (and range of Hurst exponent) the novel methods provide an unbiased estimate. The unbiasedness range for both D_α and H is wider than in the case of traditional methods, but often fails for extreme values (e.g. Hurst exponent close to zero or one).

- **consistency of the estimation**; The standard deviation of the estimate has been studied for various sample sizes, ranging from 10^3 to 10^6. To experimentally prove the consistency of the methods, it was calculated that the standard deviation significantly decreases with increasing amount of data.

The statistical testing enabled the methods to be confidently used on real data. As we had methods for many data structures, such as point sets, signals and images, this allowed to perform several case studies. The main achievements of the application parts are following:

- **detection of Alzheimer disease via EEG**; The rotational spectrum has been applied to EEG signals of control normal and Alzheimer diseased patients. Using state space reconstruction, it was possible to find significant differences in correlation dimension between the two patient groups. The patients with Alzheimer disease had smaller value of correlation dimension, which is in-line with the recent scientific findings. The precise estimation of correlation dimension has been identified as a powerful feature that carries new type of information about the investigated signal.

- **predictability analysis of stock market indices**; Revisited zero-crossing method and theory of optimal segmentation have discovered stock indices such as SMI or HSI that are less volatile in time by terms of Hurst exponent changes and have Hurst exponent value above 1/2. Therefore, these markets can be suitable for investment or prediction-making.

- **negative memory of autoluminescence of mung beans**; To investigate rather short input series of biological autoluminiscence, the novel method of fractional Brownian bridge has been applied to estimate its Hurst exponent. Via comparing with both reference signal and artificial noise it has been found out that the mung beans signal has negative memory. This indicates that the live of the mung beans is not completely random and could be connected with the nature of underlying chemical processes. The case study was published in journal PLOS ONE.

- **detection of breast cancer from mammography**; Craunio caudal mammography images were used for a detection of breast cancer. The Hurst exponent was estimated both for patients with benign and cancerous lump. The patients with cancerous lump had statistically higher value of Hurst exponent. The results were published in journal Physica A.

The often underestimated topic of fractal sets and dimension estimation has proven to be perspective in many fronts. The contribution to the current state of art has been performed both on the theoretical and application side: the work resulted in 7 published papers in journals with impact factor and 6 peer-reviewed conference papers.

One of the possible continuation of this effort could be optimization for speed. The current methods have good statistical foundations, but sometimes they can be very computationally expensive. For larger sets exceeding 10^6 points, the estimation of dimension can cost several minutes of computational time. Optimization of the current algorithms without the need of sacrificing accuracy would be a great improvement that could help to implement them into applications in industry. The second possible continuation would be to extend the applications outside the biomedical and econometric sphere. It is believed that many real world phenomena, such as load of computer networks, music generation, geology or animal behaviour have fractal character and could be modelled via the tools of fractal geometry. Moreover, the comparison of different Renyi dimension estimates for various parameters α and understanding on what kind of information they carry in a particular real-world application might be a rewarding exercise.

Bibliography

[1] H. E. Hurst, "Methods of using long-term storage in reservoirs.," *Proceedings of the Institution of Civil Engineers*, vol. 5, no. 5, pp. 519–543, 1956.

[2] B. B. Mandelbrot and J. W. V. Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[3] K. J. Falconer, *Fractal geometry : Mathematical foundations and applications.* Chichester: Wiley, 2003.

[4] W. Hurewicz, *Dimension theory.* Princeton: Princeton University Press, 2015.

[5] D. Khoshnevisan and Y. Xiao, "Lévy processes: Capacity and Hausdorff dimension," *The Annals of Probability*, vol. 33, no. 3, pp. 841–878, 2005.

[6] J. Li, Q. Du, and C. Sun, "An improved box-counting method for image fractal dimension estimation," *Pattern Recognition*, vol. 42, no. 11, pp. 2460–2469, 2009.

[7] C. Bandt, N. V. Hung, and H. Rao, "On the open set condition for self-similar fractals," *Proceedings of the American Mathematical Society*, vol. 134, no. 05, pp. 1369–1374, 2005.

[8] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical Review Letters*, vol. 50, no. 5, pp. 346–349, 1983.

[9] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D: Nonlinear Phenomena*, vol. 9, no. 1-2, pp. 189–208, 1983.

[10] Y. Wu and S. Verdu, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3748, 2010.

[11] A. Renyi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, University of California Press, 1961.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[13] R. V. L. Hartley, "Transmission of information," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928.

[14] G. M. Bosyk, M. Portesi, and A. Plastino, "Collision entropy and optimal uncertainty," *Physical Review A*, vol. 85, no. 1, 2012.

[15] J. Hutchinson, "Fractals and self similarity," *Indiana University Mathematics Journal*, vol. 30, no. 5, pp. 713–747, 1981.

[16] G. Cantor, "Ueber unendliche, lineare punktmannichfaltigkeiten," *Mathematische Annalen*, vol. 21, no. 4, pp. 545–591, 1883.

[17] E. Ott, *Chaos in dynamical systems*. Cambridge: Cambridge University Press, 1993.

[18] W. Sierpiński, "Sur une courbe *cantor*ienne qui contient une image biunivoque et continue de toute courbe donnée," *C. R. Acad. Sci., Paris*, vol. 162, pp. 629–632, 1916.

[19] H. Takayasu, *Fractals in the physical sciences*. Manchester: Manchester University Press, 1990.

[20] G. H. Hardy, "Weierstrasss non-differentiable function," *Transactions of the American Mathematical Society*, vol. 17, no. 3, p. 301, 1916.

[21] B. R. Hunt, "The Hausdorff dimension of graphs of Weierstrass functions," *Proceedings of the American Mathematical Society*, vol. 126, no. 3, pp. 791–800, 1998.

[22] B. Dubuc, "Evaluating the fractal dimension of surfaces," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 425, no. 1868, pp. 113–127, 1989.

[23] H. Whitney, "Differentiable manifolds," *The Annals of Mathematics*, vol. 37, no. 3, p. 645, 1936.

[24] H. Whitney, *Hassler Whitney collected papers*. Boston: Birkhaeuser, 1992.

[25] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, (Berlin), pp. 366–381, Springer Berlin Heidelberg, 1981.

[26] Y. Cao and S. Kiriki, "The basin of the strange attractors of some Hénon maps," *Chaos, Solitons & Fractals*, vol. 11, no. 5, pp. 729–734, 2000.

[27] M. Henon, "A two-dimensional mapping with a strange attractor," *Comm. Math. Phys.*, vol. 50, no. 1, pp. 69–77, 1976.

[28] R. Lozi, "Un attracteuré trange du type attracteur de Hénon," *Le Journal de Physique Colloques*, vol. 39, no. C5, pp. C5–9–C5–10, 1978.

[29] M. Misiurewicz and S. Štimac, "Symbolic dynamics for Lozi maps," *Nonlinearity*, vol. 29, no. 10, pp. 3031–3046, 2016.

[30] Y. Wu and S. Verdu, "Renyi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inform. Theory*, vol. 56, no. 8, pp. 3721–3748, 2010.

[31] B. C. Brookes and P. Levy, "Theorie de laddition de variables aleatoires," *The Mathematical Gazette*, vol. 39, no. 330, p. 344, 1955.

[32] B. D. Hughes, *Random walks and random environments*. New York: Clarendon Press Oxford University Press, 1995.

[33] P. Amblard, J.-F. Coeurjolly, F. Lavancier, and A. Philippe, "Basic properties of the multivariate fractional Brownian motion," *Séminaires et Congrès- SMF*, vol. 28, pp. 65–87, 2013.

[34] P.-O. Amblard and J.-F. Coeurjolly, "Identification of the multivariate fractional Brownian motion," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5152–5168, 2011.

[35] J. Unterberger, "A rough path over multidimensional fractional Brownian motion with arbitrary Hurst index by fourier normal ordering," *Stochastic Processes and their Applications*, vol. 120, no. 8, pp. 1444 – 1472, 2010.

[36] M. L. Stein, "Fast and exact simulation of fractional Brownian surfaces," *Journal of Computational and Graphical Statistics*, vol. 11, no. 3, pp. 587–599, 2002.

[37] H. Hentschel and I. Procaccia, "The infinite number of generalized dimensions of fractals and strange attractors," *Physica D: Nonlinear Phenomena*, vol. 8, no. 3, pp. 435–444, 1983.

[38] P. Grassberger, "Generalized dimensions of strange attractors," *Physics Letters A*, vol. 97, no. 6, pp. 227–230, 1983.

[39] S. B. Lowen, "Efficent generation of fractional Brownian motion for simulation of infrared focal-plane array calibration drift," *Methodology And Computing In Applied Probability*, vol. 1, no. 4, pp. 445–456, 1999.

[40] I. N. Sneddon, *Fourier Transforms (Dover Books on Mathematics)*. New York: Dover Publications, 2010.

[41] P. Abry and F. Sellan, "The wavelet-based synthesis for fractional Brownian motion proposed by f. Sellan and Y. Meyer: Remarks and fast implementation," *Applied and Computational Harmonic Analysis*, vol. 3, no. 4, pp. 377–383, 1996.

[42] Y. Meyer, F. Sellan, and M. S. Taqqu, "Wavelets, generalized white noise and fractional integration: The synthesis of fractional Brownian motion," *The Journal of Fourier Analysis and Applications*, vol. 5, no. 5, pp. 465–494, 1999.

[43] V. Pipiras, "Wavelet-based simulation of fractional Brownian motion revisited," *Applied and Computational Harmonic Analysis*, vol. 19, no. 1, pp. 49–60, 2005.

[44] R. T. Baillie, "Long memory processes and fractional integration in econometrics," *Journal of Econometrics*, vol. 73, no. 1, pp. 5–59, 1996.

[45] G. L. e. a. J-M. Bardet, "Generators of long-range dependent processes: A survey," in *Theory and applications of long-range dependence*, pp. 579–623, Boston: Birkhäuser, 2003.

[46] R. B. Davies and D. S. Harte, "Tests for Hurst effect," *Biometrika*, vol. 74, no. 1, pp. 95–101, 1987.

[47] P. F. Craigmile, "Simulating a class of stationary Gaussian processes using the Davies-Harte algorithm, with application to long memory processes," *Journal of Time Series Analysis*, vol. 24, no. 5, pp. 505–511, 2003.

[48] J.-F. Coeurjolly, "Simulation and identification of the fractional Brownian motion: a bibliographical and comparative study," *Journal of Statistical Software, Articles*, vol. 5, no. 7, pp. 1–53, 2000.

[49] A. T. A. Wood and G. Chan, "Simulation of stationary Gaussian processes in [0,1]d," *Journal of Computational and Graphical Statistics*, vol. 3, no. 4, pp. 409–432, 1994.

[50] J. Azais, "Conditions for convergence of number of crossings to the local time. Application to stable processes with independent increments and to Gaussian processes," *Probability and Mathematical Statistics*, vol. 11, no. 1, pp. 1–23, 1989.

[51] A. Feuerverger, P. Hall, and A. Wood, "Estimation of fractal index and fractal dimension of a Gaussian process by counting the number of level crossings," *Journal of Time Series Analysis*, vol. 15, pp. 587 – 606, 2008.

[52] P. Whittle, "Estimation and information in stationary time series," *Arkiv för Matematik*, vol. 2, no. 5, pp. 423–434, 1953.

[53] J. Istas and G. Lang, "Quadratic variations and estimation of the local Hölder index of a Gaussian process," *Annales de l'I.H.P. Probabilités et statistiques*, vol. 33, no. 4, pp. 407–436, 1997.

[54] L. Grafakos, *Classical Fourier Analysis: Graduate Texts in Mathematics*. New York: Springer, 2014.

[55] P. Debye, "Zerstreuung von röntgenstrahlen," *Ann. Phys.*, vol. 351, no. 6, pp. 809–823, 1915.

[56] S. Krantz, *Handbook of complex variables*. Boston: Birkhauser, 1999.

[57] M. Abramowitz, *Handbook of mathematical functions, with formulas, graphs, and mathematical tables*. New York: Dover Publications, 1965.

[58] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[59] R. Fischer, "The adaptive resolution concept in form-free distribution estimation," in *W. Kluge, Ed., Proceedings of the Workshop on Physics and Computer Science*, (Heidelberg), 1999.

[60] V. D. W. von der Linden, R. Fischer, "Adaptive kernels andOccam's razor in inversion problems," in *M. Sears, V. Nedeljkovic, N. Pendock, S. Sibisi (Eds.), Proceedings of the Maximum Entropy Conference*, (NBM printers, Port Elizabeth, South Africa), 1997.

[61] Z. Wu, "Generalized bochner's theorem for radial function," *Approximation Theory and its Applications*, vol. 13, pp. 47–57, 1997.

[62] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[63] A. Gelman, *Bayesian data analysis*. Boca Raton: CRC Press, 2014.

[64] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[65] W. Perks, "Some observations on inverse probability including a new indifference rule," *Journal of the Institute of Actuaries (1886-1994)*, vol. 73, no. 2, pp. 285–334, 1947.

[66] J. Luo, Z. Xie, and M. Xie, "Interpolated DFT algorithms with zero padding for classic windows," *Mechanical Systems and Signal Processing*, vol. 70-71, pp. 1011–1025, 2016.

[67] B. R. Jefferies, *Spectral Properties of Noncommuting Operators (Lecture Notes in Mathematics)*. Berlin: Springer, 2004.

[68] G. Seber and C. Wild, *Nonlinear regression*. Hoboken: Wiley, 2003.

[69] J. Wishart, "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population," *Biometrika*, vol. 20A, no. 1-2, pp. 32–52, 1928.

[70] T. Hida and M. Hitsuda, *Gaussian Processes (Translations of Mathematical Monographs)*. Providence: American Mathematical Society, 2007.

[71] S. R. Eliason, *Maximum Likelihood Estimation: Logic And Practice (Quantitative Applications in the Social Sciences)*. Newbury Park: SAGE Publications, Inc, 1993.

[72] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan, "Analysis of circulant embedding methods for sampling stationary random fields," *SIAM Journal on Numerical Analysis*, vol. 56, no. 3, pp. 1871–1895, 2018.

[73] C. R. Dietrich and G. N. Newsam, "Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix," *SIAM Journal on Scientific Computing*, vol. 18, no. 4, pp. 1088–1107, 1997.

[74] J. Hoffman, *Numerical methods for engineers and scientists*. New York: Marcel Dekker, 2001.

[75] R. Gonzalez, *Digital Image processing using MATLAB*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2004.

[76] N. Wiener, *Time series*. Cambridge: The MIT Press, 1975.

[77] G. Edgar, *Measure, topology, and fractal geometry*. New York: Springer-Verlag, 2008.

[78] D. A. Russell, J. D. Hanson, and E. Ott, "Dimension of strange attractors," *Phys. Rev. Lett.*, vol. 45, pp. 1175–1178, 1980.

[79] H. Peitgen, *Chaos and fractals : New frontiers of science*. New York: Springer-Verlag, 1992.

[80] G. Chan and A. T. A. Wood, "Algorithm as 312: An algorithm for simulating stationary Gaussian random fields," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 46, no. 1, pp. 171–181, 1997.

[81] S. Suman, A. Kumar, and G. K. Singh, "A new method for higher-order linear phase FIR digital filter using shifted Chebyshev polynomials," *Signal, Image and Video Processing*, vol. 10, no. 6, pp. 1041–1048, 2015.

[82] P. Mohindru, R. Khanna, and S. S. Bhatia, "New tuning model for rectangular windowed fir filter using fractional Fourier transform," *Signal, Image and Video Processing*, vol. 9, no. 4, pp. 761–767, 2013.

[83] S. Salivahanan, *Digital signal processing*. New Delhi: Tata McGraw Hill Education Pvt. Ltd, 2010.

[84] F. J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3/4, p. 246, 1948.

[85] G. Casella and R. L. Berger, *Statistical Inference*. Australia Pacific Grove: Cengage Learning, 2001.

[86] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," *Proceedings of the Fourth International Workshop on Digital Mammography*, 2000.

[87] M. D. Heath, K. Bowyer, D. Kopans, W. P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography / IWDM*, 1998.

[88] C. Feng, Y. Mei, and B. Vidakovic, *Mammogram diagnostics Using robust wavelet-based estimator of Hurst Exponent*, pp. 109–140. Cham: Springer International Publishing, 2018.

Attachments

Publications in impact journals

M. Dlask and J. Kukal, "Application of rotational spectrum for correlation dimension estimation", *Chaos, Solitons and Fractals*, vol. 99, pp. 256-262, 2017.

M. Dlask and J. Kukal, "Hurst exponent estimation from short time series", *Signal, Image and Video Processing*, vol. 13, pp. 263-269, 2018.

M. Dlask and J. Kukal, "Translation and rotation invariant method of Renyi dimension estimation", *Chaos, Solitons & Fractals*, vol. 114, pp. 536-541, 2018.

M. Dlask and J. Kukal, "Alzheimer disease diagnostics from EEG via Wishart distribution of fractional processes", *Signal, Image and Videoprocessing*, vol. 15, pp. 1435-1442, 2021.

M. Dlask and J. Kukal, "Hurst exponent estimation of fractional surfaces for mammogram images analysis", *Physica A*, vol. 585, pp. 126424, 2022.

M. Dlask, J. Kukal, M. Poplova, P. Sovka and M. Cifra, "Short-time fractal analysis of biological autoluminescence", *PLOS ONE*, vol. 14, pp. 1-17, 2019.

M. Dlask, J. Kukal and O. Vysata, "Bayesian approach to Hurst exponent estimation", *Methodology and Computing in Applied Probability*, vol. 19, pp. 973-983, 2017.

Peer-reviewed publications in conference proceedings

M. Dlask and J. Kukal, "Correlation dimension as a measure of stock market variability", *Proceedings of 35th International Conference Mathematical Methods in Economics*, pp. 119-124, 2017.

M. Dlask and J. Kukal. "Correlation Dimension Estimation from EEG Time Series for Alzheimer Disease Diagnostics", *Proceedings of the International Conference on Bioinformatics Research and Applications*, pp. 62-65, 2017.

M. Dlask, J. Kukal and P. Sovka. "Fractional Brownian bridge model for Alzheimer disease detection from EEG signal", *Proceedings of International Conference on Signal Processing and Information Security*, pp. 1-4, 2018.

M. Dlask, J. Kukal, and Q.V. Tran, "Revisited Zero- Crossing Method for Hurst Exponent Estimation in Time Series", *Proceedings of International Conference Mathematical Methods in Economics*, pp. 115-120, 2015.

M. Dlask, "Fractional Brownian bridge as a tool for short time series analysis", *Proceedings of International Conference Mathematical Methods in Economics*, pp. 149-154, 2016.

L. Tylova, M. Dlask, J. Kukal, and Q.V. Tran, "Could prediction error help in fractal analysis of time series?", *Proceedings of International Conference Mathematical Methods in Economics*, pp. 847-852, 2015.

Other publications and talks

M. Dlask, "Novel approach of fractal dimension estimation based on optimal segmentation of time series", *Workshop on Scientific computing*, Decin, 2015.

M. Dlask, "On the correlation dimension estimation using rotational spectrum, *Workshop of scientific computing*, Decin, 2016.

M. Dlask, "On the investigation of Julia sets using rotational spectrum", *Doktorandske dny*, Prague, 2016.

M. Dlask, "Application of correlation dimension for Alzheimer disease diagnostics", *Workshop on Scientific computing*, Decin, 2017.

M. Dlask, "Bayesian approach to Hurst exponent estimation", *Doktorandske dny*, Prague, 2017.

M. Dlask, "Translation and Rotation Invariant Method of Renyi Dimension Estimation", *Doktorandske dny*, Prague, 2018.

M. Dlask, "Rotational spectrum as a tool for Alzheimer disease diagnostics", *Workshop on Scientific computing*, Decin, 2019.

M. Dlask, "Short-time fractional analysis of biological autoluminescence", *Doktorandske dny*, Prague, 2019.

M. Dlask, "Maximum likelihood approach to Hurst exponent estimation of fractional time series", *Doktorandske dny*, Prague, 2020.

M. Dlask, "Application of multidimensional fBm in mammography screening", *Workshop on Scientific computing*, Decin, 2021.

Academic Experience Abroad

Conferences

ICBRA 2017, 2017 International Conference on Bioinformatics Research and Applications, Barcelona, Spain, 8th December 2017 - 10th December 2017

ICSPIS 2018, 2018 International Conference on Signal Processing and Information Security, Dubai, United Arab Emirates, 7th November - 8th November 2018

Study stays related to applied statistics

Telecom ParisTech, Multimedia Indexing and Retrieval, Paris, France, 13th March - 21st March 2015

University of Jyväskylä, 25th Jyväskylä Summer School, Jyväskylä, Finland, 5th August - 21st August 2015

The Hong Kong University of Science and Technology, Summer Camp for Elite Students on Computer Science and Engineering department, Hong Kong, 3rd July - 9th July 2016

Tor Vergata University of Rome, The Big Data Theory Course, Rome, Italy, 25 July - 31st July 2016

Other study stays abroad

University Las Palmas de Gran Canaria, Las Palmas Winter Course 2016, Las Palmas, Gran Canaria, Spain, 31st January - 11th February 2017

Istanbul Technical University, Sustainable Built Environment and Sustainability Management, Istanbul, Turkey, 11th March - 19th March 2017

Konkuk University, 2017 KU International Summer Program, Seoul, South Korea, 2nd July - 15th July 2017

Selected Papers Published in Impact Journals

During the PhD study, there have been published seven papers in impact journals. This attachment presents four most important papers in their original form as accepted to journals:

- M. Dlask and J. Kukal, "Application of rotational spectrum for correlation dimension estimation", *Chaos, Solitons and Fractals*, vol. 99, pp. 256-262, 2017.

- M. Dlask and J. Kukal, "Translation and rotation invariant method of Renyi dimension estimation", *Chaos, Solitons & Fractals*, vol. 114, pp. 536-541, 2018.

- M. Dlask and J. Kukal, "Hurst exponent estimation of fractional surfaces for mammogram images analysis", *Physica A*, vol. 585, pp. 126424, 2022.

- M. Dlask, J. Kukal, M. Poplova, P. Sovka and M. Cifra, "Short-time fractal analysis of biological autoluminescence", *PLOS ONE*, vol. 14, pp. 1-17, 2019.

Chaos, Solitons and Fractals 99 (2017) 256–262

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Application of rotational spectrum for correlation dimension estimation

Martin Dlask*, Jaromir Kukal

Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Department of Software Engineering, Trojanova 12, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 22 May 2016
Revised 15 January 2017
Accepted 12 April 2017

MSC:
28A80
65P20

Keywords:
Point set
Correlation dimension
Power spectrum
Rotation
Monte Carlo

ABSTRACT

Correlation dimension is one of the many types of fractal dimension. It is usually estimated from a finite number of points from a fractal set using correlation sum and regression in a log-log plot. However, this traditional approach requires a large amount of data and often leads to a biased estimate. The novel approach proposed here can be used for the estimation of the correlation dimension in a frequency domain using the power spectrum of the investigated fractal set. This work presents a new spectral characteristic called "rotational spectrum" and shows its properties in relation to the correlation dimension. The theoretical results can be directly applied to uniformly distributed samples from a given point set. The efficiency of the proposed method was tested on sets with a known correlation dimension using Monte Carlo simulation. The simulation results showed that this method can provide an unbiased estimation for many types of fractal sets.

1. Introduction

Correlation dimension D_2 is a popular tool for fractal dimension estimation and belongs to a family of entropy-based fractal dimensions such as capacity dimension D_0, information dimension D_1 and their generalisation, Renyi dimension D_α, for $\alpha \geq 0$. The properties of the different dimension types are summarised in [1] and [2]. The main idea of using correlation dimension is the distance between its points in space. In the original concept, only the number of points that are not farther apart as a fixed value can carry the information about the density of points contained in the investigated set. The geometrical meaning of correlation dimension is explained well in [3].

This traditional approach of correlation dimension estimation is based on Grassberger and Procaccia's algorithm [4,5] and is widely used in biomedicine for electroencephalography signal analysis [6,7] or in cardiology [8]. Recently, new approaches of correlation dimension estimation were presented using a weighting function [9] or methods suitable for high-dimensional signals [10]. The linear regression model, on which the majority of methods are based, provides an often biased estimate of fractal dimension; for this reason, Hongying and Duanfeng [11] made some efforts to improve this procedure.

In this work, we present a novel approach of correlation dimension estimation that is based on the rotation of the power spectrum of a point set. The proposed method is stable even for a small number of points, and the resulting characteristic has a smooth development.

2. Correlation dimension

Correlation dimension, introduced by Grassberger and Procaccia, involves measuring the distance between all pairs of points in the investigated set. For the Lebesgue measurable set $\mathcal{F} \subset \mathbb{R}^n$, the *correlation sum* [4] is defined for $r > 0$ as the limit case

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I(\|\boldsymbol{x_i} - \boldsymbol{x_j}\| \leq r), \tag{1}$$

where $\|\cdot\|$ denotes a Euclidean norm that is rotation invariant, I is the indicator function and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are vectors from $\mathcal{F}$. Because the correlation dimension expresses the relative amount of points whose distance is less than r, the correlation sum can be rewritten as

$$C(r) = \mathop{E}_{\boldsymbol{x},\boldsymbol{y}\sim U(\mathcal{F})} I(\|\mathbf{x} - \mathbf{y}\| \leq r) = \mathop{\text{prob}}_{\boldsymbol{x},\boldsymbol{y}\sim U(\mathcal{F})} (\|\boldsymbol{x} - \boldsymbol{y}\| \leq r), \tag{2}$$

for $\boldsymbol{x}, \boldsymbol{y}$ that are uniformly distributed on $\mathcal{F}$. Therefore, $C(r)$ is a cumulative distribution function of random variable $r = \|\boldsymbol{x} - \boldsymbol{y}\|$. The

* Corresponding author.
E-mail address: martindlask@centrum.cz (M. Dlask).

http://dx.doi.org/10.1016/j.chaos.2017.04.026
0960-0779/

correlation dimension D_2 of set $\mathcal{F}$ is based on the correlation sum and is defined as

$$D_2 = \lim_{r\to 0^+} \frac{\ln \mathrm{C}(r)}{\ln r}, \tag{3}$$

if the limit exists.

3. Continuous spectrum of a point set

The Fourier transform of an n-dimensional set $\mathcal{F} \subset \mathbb{R}^n$ is defined using the operator of the expected value [12] as

$$\mathrm{F}(\boldsymbol{\omega}) = \mathop{\mathrm{E}}_{\boldsymbol{x}\sim \mathrm{U}(\mathcal{F})} \exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{x}) \tag{4}$$

for angular frequency $\boldsymbol{\omega} \in \mathbb{R}^n$ and for $\boldsymbol{x}$ uniformly distributed on $\mathcal{F}$. The *power spectrum* of set $\mathcal{F}$ equals $\mathrm{P}(\boldsymbol{\omega}) = |\mathrm{F}(\boldsymbol{\omega})|^2 = \mathrm{F}(\boldsymbol{\omega})\cdot \mathrm{F}^*(\boldsymbol{\omega})$, where F^* is a complex conjugate of F. Moreover, it can be expressed as

$$\begin{aligned}\mathrm{P}(\boldsymbol{\omega}) &= \mathop{\mathrm{E}}_{\boldsymbol{x}\sim \mathrm{U}(\mathcal{F})} \mathop{\mathrm{E}}_{\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{x}) \exp(\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{y}) \\ &= \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \exp(-\mathrm{i}\boldsymbol{\omega}\cdot(\boldsymbol{x}-\boldsymbol{y})),\end{aligned} \tag{5}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent and identically distributed from $\mathcal{F}$. The power spectrum is frequently used for fractal set investigation [13–15]. When the research is physically motivated, it is usual to denote the angular frequency as $\omega = 2\pi/\lambda$ for wavelength λ of an X-ray or light beam.

4. Rotational spectrum

The goal of the novel method is to obtain a one-dimensional function as a derivative of the power spectrum, which is useful in fractal analysis. The procedure was inspired by Debye [16] and by his X-ray diffraction method, which is often referred to as the Debye-Scherrer method. We denote $\mathrm{SO}(n)$ as the group of all rotations in $\mathbb{R}^n$ around the origin. Because any rotation $\mathrm{R} \in \mathrm{SO}(n)$ is a linear transform, the following equation holds

$$\mathrm{R}(\boldsymbol{x}) - \mathrm{R}(\boldsymbol{y}) = \mathrm{R}(\boldsymbol{x}-\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|\cdot\boldsymbol{\xi}, \tag{6}$$

where $\boldsymbol{\xi}$ is a direction vector satisfying $\|\boldsymbol{\xi}\| = 1$ and $\boldsymbol{\xi} \in \mathcal{S}_{n-1}$ for an n-dimensional sphere $\mathcal{S}_{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$. Using the factorisation of angular frequency $\boldsymbol{\omega} = \Omega\cdot\boldsymbol{\psi}$ for $\Omega \in \mathbb{R}_0^+$ and normalisation vector $\boldsymbol{\psi} \in \mathcal{S}_{n-1}$, we can define *rotational spectrum* as

$$\mathrm{S}(\Omega) = \mathop{\mathrm{E}}_{\mathrm{R}\in \mathrm{SO}(n)} \mathop{\mathrm{E}}_{\boldsymbol{\psi}\in\mathcal{S}_{n-1}} \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \exp(-\mathrm{i}\Omega\boldsymbol{\psi}\mathrm{R}(\boldsymbol{x}-\boldsymbol{y})), \tag{7}$$

which can be expressed explicitly in the following theorem.

Theorem 1. *Rotational spectrum can be expressed as*

$$\mathrm{S}(\Omega) = \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \mathrm{H}_n(\Omega\|\boldsymbol{x}-\boldsymbol{y}\|), \tag{8}$$

where

$$\mathrm{H}_n(q) = \frac{2^{\frac{n-2}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}{q^{\frac{n-2}{2}}}\mathrm{J}_{\frac{n-2}{2}}(q). \tag{9}$$

Proof. Because every rotation is a linear transform, we can rewrite the rotational spectrum as

$$\mathrm{S}(\Omega) = \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \mathop{\mathrm{E}}_{\boldsymbol{\psi},\boldsymbol{\xi}\in\mathcal{S}_{n-1}} \exp(-\mathrm{i}\Omega\|\boldsymbol{x}-\boldsymbol{y}\|\boldsymbol{\psi}\cdot\boldsymbol{\xi}). \tag{10}$$

The angle ν between vectors $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ satisfies $\cos\nu = \boldsymbol{\psi}\cdot\boldsymbol{\xi}$. Without loss of generality, we can set $\boldsymbol{\xi} = (1, 0, 0, \ldots, 0)$ and rewrite the rotational spectrum as

$$S(\Omega) = \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{F}} H_n(\Omega\|\boldsymbol{x}-\boldsymbol{y}\|), \tag{11}$$

where the function $\mathrm{H}_n : \mathbb{R} \mapsto \mathbb{C}$ is defined as

$$\mathrm{H}_n(q) = \mathop{\mathrm{E}}_{\substack{\boldsymbol{\psi}\in\mathcal{S}_{n-1}\\ \psi_1=\cos\nu}} \exp(-\mathrm{i}q\cos\nu). \tag{12}$$

For $n = 1$, we obtain a degenerated rotation together with $\nu \in \{0, \pi\}$; therefore, the kernel function H_1 equals

$$\mathrm{H}_1(q) = \frac{\exp(-\mathrm{i}q) + \exp(\mathrm{i}q)}{2} = \cos q. \tag{13}$$

In case $n \geq 2$, we can express the kernel function using an integral formula:

$$\mathrm{H}_n(q) = \frac{\mathrm{I}_1(q)}{\mathrm{I}_2(q)} = \frac{\int_0^\pi \exp(-\mathrm{i}q\cos\nu)\sin^{n-2}\nu\,\mathrm{d}\nu}{\int_0^\pi \sin^{n-2}\nu\,\mathrm{d}\nu}. \tag{14}$$

The Poisson integral [17] formula for the Bessel function $\mathrm{J}_p(q)$ of the first kind in the form

$$\mathrm{J}_p(q) = \frac{\left(\frac{q}{2}\right)^p}{\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}\int_0^\pi \exp(-\mathrm{i}q\cos\nu)\sin^{2p}\nu\,\mathrm{d}\nu \tag{15}$$

allows the integral in the nominator to be rewritten as

$$\mathrm{I}_1(q) = \frac{\mathrm{J}_p(q)\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\left(\frac{q}{2}\right)^p}, \tag{16}$$

whereas the integral in the denominator is a limit case of the Poisson formula

$$\mathrm{I}_2(q) = \lim_{q\to 0} \frac{\mathrm{J}_p(q)\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\left(\frac{q}{2}\right)^p} = \frac{\Gamma\left(p+\frac{1}{2}\right)\sqrt{\pi}}{\Gamma(p+1)}. \tag{17}$$

For $p = \frac{n-2}{2}$, we obtain the final form of the kernel function expressed by the Bessel function $\mathrm{J}_p(q)$ as

$$\mathrm{H}_n(q) = \frac{2^{\frac{n-2}{2}}\cdot\Gamma\left(\frac{n}{2}\right)}{q^{\frac{n-2}{2}}}\mathrm{J}_{\frac{n-2}{2}}(q). \tag{18}$$

Applying $\mathrm{H}_n(q)$ for $n = 1$, we obtain $\mathrm{H}_1(q) = \cos q$ as a particular case, which extends the range of formula (18) to $n \in \mathbb{R}$. □

The rotation can be performed in any space whose dimension n is not less than the dimension m of the original space of $\mathcal{F}$. When the dimension of the rotation is greater than m, any vector $\boldsymbol{x} \in \mathcal{F}$ is completed, with the zeros for the remaining $n-m$ coordinates having a sufficient length. The most valuable result can be obtained in the case of rotation in an infinite-dimensional space.

Theorem 2. *The scaled limit case of the kernel function* H_n *is the Gaussian function, i.e.,*

$$\lim_{n\to\infty} \mathrm{H}_n(t\sqrt{n}) = \exp\left(-\frac{t^2}{2}\right). \tag{19}$$

Proof. For the investigation of the behaviour of the kernel function when $n \to \infty$, we use the Taylor expansion of $H_n(q)$ centred at $q_0 = 0$

$$\mathrm{H}_n(q) = \sum_{k=0}^{\infty} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2}+k)k!}\left(-\frac{q^2}{4}\right)^k, \tag{20}$$

and by using the substitution $q = t\sqrt{n}$, we can transform it into

$$\mathrm{H}_n(t\sqrt{n}) = \sum_{k=0}^{\infty} \frac{1}{k!}\left(-\frac{t^2}{2}\right)^k \frac{\Gamma(\frac{n}{2})n^k}{\Gamma(\frac{n}{2}+k)2^k}. \tag{21}$$

For every $k \in \mathbf{N}$, it holds that

$$\lim_{n\to\infty} \frac{\Gamma(\frac{n}{2})n^k}{\Gamma(\frac{n}{2}+k)2^k} = 1, \tag{22}$$

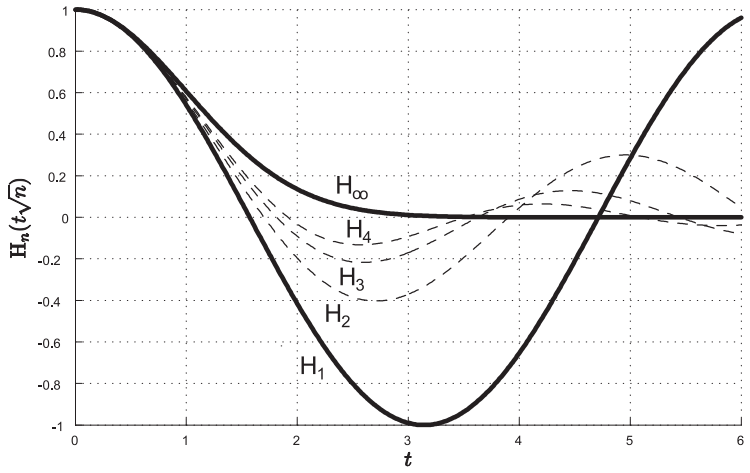


Fig. 1. Kernel functions of a rotational spectrum.

and, therefore, the limit case of the kernel function equals

$$\lim_{n\to\infty} \mathrm{H}_n(t\sqrt{n}) = \exp\left(-\frac{t^2}{2}\right). \tag{23}$$

□

For simplicity, we will use the following notation in the subsequent sections:

$$\mathrm{H}_\infty(q) = \exp\left(-\frac{q^2}{2}\right). \tag{24}$$

This type of Gaussian kernel has been widely applied in the Parzen density estimates [18–20] of a probability density function and its properties, but with another meaning and motivation. The behaviour of the $\mathrm{H}_n(q)$ kernels is visualised in Fig. 1 for various dimensions.

5. Relationship to correlation dimension

In this section, we discuss the relationship between the rotational spectrum for the limit kernel H_∞ and the correlation dimension. The correlation sum is a cumulative distribution function of the distances between the points in a fractal set; therefore, the rotational spectrum can be written as a Stieltjes integral:

$$\mathrm{S}(\Omega) = \int_0^\infty \mathrm{H}_\infty(\Omega r)\mathrm{dC}(r) = \int_0^\infty \exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{dC}(r). \tag{25}$$

After the application of the integration by parts, we can obtain

$$\mathrm{S}(\Omega) = \int_0^\infty \Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{C}(r)\Omega \mathrm{d}r, \tag{26}$$

and by substituting $\xi = \Omega r$, we get the integral formula for the rotational spectrum:

$$\mathrm{S}(\Omega) = \int_0^\infty \xi \cdot \mathrm{C}\left(\frac{\xi}{\Omega}\right)\exp\left(-\frac{\xi^2}{2}\right)\mathrm{d}\xi. \tag{27}$$

Theorem 3. *Let $\mathcal{F} \subset \mathbb{R}^n$ be a Lebesgue measurable set with the rotational spectrum*

$$\mathrm{S}(\Omega) = \underset{\boldsymbol{x},\boldsymbol{y}\sim\mathrm{U}(\mathcal{F})}{E} \mathrm{H}_\infty(\Omega\|\boldsymbol{x}-\boldsymbol{y}\|), \tag{28}$$

and let us assume that correlation dimension D_2 (3) exists. Then, it holds that

$$\lim_{\Omega\to\infty} \frac{\ln \mathrm{S}(\Omega)}{\ln \Omega} = -D_2. \tag{29}$$

Proof. To prove this, let us suppose that $\delta < 1$ and that, at first, $r < \delta$. Owing to the existence of correlation dimension, we have $\forall\epsilon > 0\, \exists\delta > 0$

$$0 < r < \delta \Rightarrow \left|\frac{\ln \mathrm{C}(r)}{\ln r} - D_2\right| < \epsilon,$$

and, therefore,

$$r^{D_2+\epsilon} < \mathrm{C}(r) < r^{D_2-\epsilon}. \tag{30}$$

However, for $r \geq \delta$, we have

$$\delta^{D_2+\epsilon} < C(r) \leq 1. \tag{31}$$

Now, we can estimate the lower and the upper boundary for the spectrum

$$\begin{aligned} S(\Omega) &= \underset{\boldsymbol{x},\boldsymbol{y}\sim\mathrm{U}(\mathcal{F})}{E} \exp\left(-\Omega^2\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2}\right) \\ &= \int_0^\infty C(r)\Omega^2 r\exp\left(-\Omega^2\frac{r^2}{2}\right)\mathrm{d}r \end{aligned} \tag{32}$$

as

$$\mathrm{I}_\mathrm{L}(\Omega) < \mathrm{S}(\Omega) < \mathrm{I}_U(\Omega). \tag{33}$$

We can rewrite I_U as

$$\begin{aligned} \mathrm{I}_U(\Omega) = &\int_0^\delta r^{D_2-\epsilon}\Omega^2 r\exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{d}r \\ &+ \int_\delta^\infty \Omega^2 r\exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{d}r, \end{aligned} \tag{34}$$

and after the substitution $t = \Omega^2 r^2/2$, we get

$$\mathrm{I}_U(\Omega) = \Omega^{\epsilon-D_2}\cdot 2^{\frac{D_2-\epsilon}{2}}\cdot\int_0^{\Omega^2\delta^2/2} t^{\frac{D_2-\epsilon}{2}}\exp(-t)\mathrm{d}t + \exp\left(\frac{-\Omega^2\delta^2}{2}\right). \tag{35}$$

Therefore, the upper bound I_U can be expressed as

$$\mathrm{I}_U(\Omega) = \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2-\epsilon}\cdot\frac{D_2-\epsilon}{2}\cdot\Gamma_{\mathrm{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2-\epsilon}{2}\right), \tag{36}$$

where Γ_{inc} is an incomplete Gamma function. It is possible to do an estimation from above as

$$\mathrm{I}_U < \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2-\epsilon}\cdot\frac{D_2-\epsilon}{2}\cdot\Gamma\left(\frac{D_2-\epsilon}{2}\right). \tag{37}$$

The lower bound I_L is rewritten as

$$I_L(\Omega) = \int_0^{\delta} r^{D_2+\epsilon}\Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{d}r + \int_{\delta}^{\infty} r^{D_2-\epsilon}\Omega^2 r \exp\left(-\frac{\Omega^2 r^2}{2}\right)\mathrm{d}r \tag{38}$$

and can be estimated as

$$I_L(\Omega) > \left(\frac{\sqrt{2}}{\Omega}\right)^{D_2+\epsilon} \cdot \Gamma_{\mathrm{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2+\epsilon}{2}+1\right). \tag{39}$$

Altogether, we receive the upper and the lower boundary for the logarithm of the rotational spectrum

$$(D_2-\epsilon)\left(\frac{1}{2}\ln 2 - \ln\Omega\right) + L_1(\Omega) > \ln S(\Omega) > (D_2+\epsilon)\left(\frac{1}{2}\ln 2 - \ln\Omega\right) + L_2(\Omega) \tag{40}$$

and after the rearrangement

$$-\epsilon + \frac{L_2(\Omega) + \frac{D_2+\epsilon}{2}\ln 2}{\ln\Omega} < \frac{S(\Omega)}{\ln\Omega} + D_2 < \epsilon + \frac{L_1(\Omega) + \frac{D_2-\epsilon}{2}\ln 2}{\ln\Omega} \tag{41}$$

for the functions

$$L_1(\Omega) = \ln\Gamma\left(\frac{D_2-\epsilon}{2}\right) + \ln\frac{D_2-\epsilon}{2} \tag{42}$$

and

$$L_2(\Omega) = \ln\Gamma_{\mathrm{inc}}\left(\frac{\Omega^2\delta^2}{2}, \frac{D_2+\epsilon}{2}+1\right). \tag{43}$$

It holds that both L_1 and L_2 are constrained functions of Ω. Therefore, Ω_0 exists, which guarantees that, for any $\Omega > \Omega_0 > 1$, it is valid that

$$\left|\frac{\ln S(\Omega)}{\ln\Omega} + D_2\right| < 2\epsilon = \epsilon^*, \tag{44}$$

which completes the proof. □

The Lebesgue measurability of the investigated set is an important prerequisite because it ensures the capability to perform a uniform sampling. As a general remark, we could consider another kernel function instead of H_∞. For any non-increasing function $\Phi : \mathbb{R}_0^+ \mapsto [0;1]$ satisfying $\Phi(0) = 1$ and $\Phi(\infty) = 0$, and whose first derivative $\Phi'(\xi)$ exists for any $\xi > 0$, we consider the rotational spectrum in a more general form as

$$S(\Omega) = \mathop{\mathrm{E}}_{\boldsymbol{x},\boldsymbol{y}\sim \mathrm{U}(\mathcal{F})} \Phi(\Omega\|\boldsymbol{x}-\boldsymbol{y}\|). \tag{45}$$

The Ψ function is defined as

$$\Psi(\alpha) = -\int_0^{\infty} \xi^{\alpha}\Phi'(\xi))\mathrm{d}\xi, \tag{46}$$

and the existence of limit (29) is guaranteed only if both $\Psi(D_2+\epsilon)$ and $\Psi(D_2-\epsilon)$ are finite for arbitrary $\epsilon \in (0;\epsilon_0)$s. Another example of a kernel function could be the generalised exponential kernel

$$\Phi_1(\xi) = \exp\left(-\frac{\xi^{\beta}}{\beta}\right) \tag{47}$$

for $\beta > 0$ or the inverse polynomial kernel

$$\Phi_2(\xi) = \frac{1}{P(\xi)}, \tag{48}$$

where $P(\xi)$ represents a polynomial of order $M > D_2 + 1$.

6. Method of estimation

The simulation of the rotational spectrum is based on generating point pairs using a Monte Carlo approach. The points are independently and uniformly sampled from the analysed set $\mathcal{F}$. With $M \in \mathbb{N}$ fixed and $\boldsymbol{x}_i, \boldsymbol{y}_i \sim \mathrm{U}(\mathcal{F})$, the rotational spectrum is estimated as

$$\widehat{S}(\Omega) = \frac{1}{M}\sum_{j=1}^{M} H_\infty(\Omega\|\boldsymbol{x}_j - \boldsymbol{y}_j\|) \tag{49}$$

including the variance estimate

$$\widehat{\mathrm{var}\,S}(\Omega) = \frac{1}{M-1}\sum_{j=1}^{M}(H_\infty(\Omega\|\boldsymbol{x}_j-\boldsymbol{y}_j\|) - \widehat{S}(\Omega))^2 \tag{50}$$

To take advantage of the linear dependence between the logarithm of the rotational spectrum and the logarithm of the distance, we can reasonably consider the model

$$\ln S(\Omega) = A - D_2 \cdot \ln\,\Omega + \epsilon. \tag{51}$$

The estimation of parameter D_2 is based on the maximum likelihood method using L_p regression with a minimisation criterion

$$CRIT = \sum_{k=1}^{N} |y_k - f(x_k, \boldsymbol{a})|^p \tag{52}$$

for $p > 1$ and a general model formulated as $y = f(x_k, \boldsymbol{a})$. In our case, the minimisation criterion satisfies

$$CRIT^* = \sum_{k=1}^{N} \left|\ln\,\widehat{S}(\Omega_k) - A + D_2 \ln\,\Omega_k\right|^p. \tag{53}$$

The algorithm is based on the capability to generate point pairs uniformly from a fractal set and can be formulated as follows:

- The parameter M is chosen arbitrarily, but is large enough (e.g., $M = 10^5$). This parameter represents the number of Monte Carlo simulations, which is equal to the number of point pairs from the fractal set used for the estimation.
- The values of Ω, in which the calculation is performed, are determined. For the simulation, it is recommended to choose regular sampling from the interval, where the rotational spectrum is expected to have a linear characteristic.
- The calculation of the rotational spectrum is performed at points $\Omega_1, \Omega_2 \ldots, \Omega_N$, according to Eq. (49).
- With the values of Ω_i and the respective $\widehat{S}(\Omega_i)$ estimates, it is possible to perform minimisation using Eq. (53) with a maximum likelihood method.
- The resulting parameter D_2 represents the estimate of the correlation dimension.

7. Application to simulated data

The main feature of the proposed methodology is its smoother dependence of the spectrum on Ω. We tested this property on point sets with well-known Hausdorff dimension, which are summarised in Table 1. On the left side of Fig. 2, there is a traditional log-log plot, where the logarithm of the correlation sum is plotted against the logarithm of the distances. The lines represent the upper and the lower theoretical bounds for $C(r)$. The right side of the figure shows the dependence of the logarithm of the infinite-dimensional rotational spectrum on the logarithm of the frequency (Ω). The experiment was performed on a two-dimensional Cantor dust with the contraction coefficient $a = 1/3$ and $M = 10^5$ pairs of points.

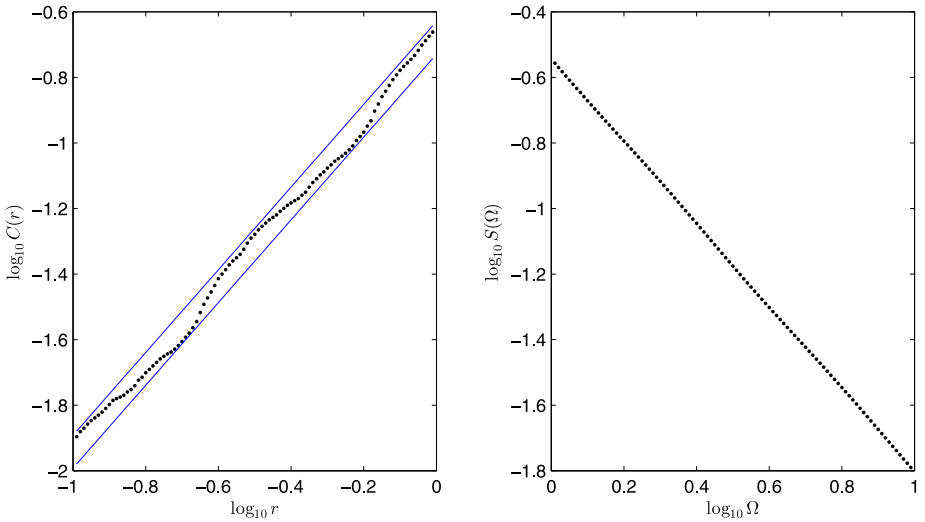


Fig. 2. Sample from the two-dimensional Cantor dust: correlation sum (left), rotational spectrum (right).

Table 1
List of point sets.

Structure	n	Parameter	Value	D_2
Cantor dust [21]	1–3	a	1/20–1/2	$-n \ln 2 / \ln a$
Levy flight trajectory [22]	2–4	α	0.2–1.0	$\min(\alpha, n)$
fBm trajectory [22]	3–4	H	0.25–0.75	$\min\left(\frac{1}{H}, n\right)$
fBm graph [23]	1	H	0.1–0.9	$2 - H$
Takagi function graph [24]	1	a	0.55–0.95	$2 + \log_2 a$

Table 2
Cantor dust analysis using linear least squares fitting.

M	Correlation sum			Rotational spectrum		
	$\widehat{D_2}$	sd	p-value	$\widehat{D_2}$	sd	p-value
10^3	1.2254	0.0648	0.2868	1.2501	0.0323	0.3579
10^4	1.2392	0.0202	0.1310	1.2689	0.0183	0.3502
10^5	1.2513	0.0039	0.0034	1.2592	0.0030	0.1915
10^6	1.2599	0.0005	$4.44 \cdot 10^{-5}$	1.2601	0.0003	$1.54 \cdot 10^{-7}$

Table 3
Cantor dust analysis using L_4.

M	Correlation sum			Rotational spectrum		
	$\widehat{D_2}$	sd	p-value	$\widehat{D_2}$	sd	p-value
10^3	1.2941	0.1178	0.3922	1.2378	0.1010	0.4059
10^4	1.2937	0.0803	0.3459	1.3019	0.0470	0.1971
10^5	1.2341	0.0574	0.3143	1.2618	0.0100	0.4976
10^6	1.2654	0.0474	0.4702	1.2609	0.0076	0.4498

The reason for the smooth development of the rotational spectrum characteristic is the infinite-dimensional rotation. It is possible to compare the correlation dimension estimate from the rotational spectrum approach and the traditional correlation sum. At first, the linear regression with least squares minimisation criterion was used to fit the model. However, the results were biased for a larger number of data points, as can be seen from Table 2. To avoid the bias, we decided to use L_p regression for the rotational spectrum fitting using a maximum likelihood method. The numerical experiments proved that any order $p \geq 4$ is appropriate to fit the model. Therefore, we considered L_4 regression for the estimation of the correlation dimension. Table 3 shows the results for different numbers of point pairs M. The estimates of $\widehat{D_2}$ based on L_4 regression were unbiased for both the correlation sum and the rotational spectrum. However, the variance of spectrum-based estimates rapidly decreased with M.

It is also possible to estimate the rotational spectrum for finite rotation using the kernel functions H_n for $n \in \mathbb{N}$. The comparison of the kernel functions that can be used for the rotation of the power spectrum is shown in Fig. 3 for H_2, H_3, H_4 and H_∞. The traditional Sierpinski carpet was used for this simulation.

Table 4
Sierpinski carpet analysis for the different kernel functions.

Kernel function	D_2	$\widehat{D_2}$	sd	p-value	$f_{\min}$	$f_{\max}$
H_2	1.8928	1.9851	0.2625	0.4106	1.0	3.0
H_4	1.8928	1.8673	0.1128	0.3624	1.0	3.0
H_7	1.8928	1.9148	0.0863	0.3993	1.0	3.0
H_{10}	1.8928	1.9019	0.0636	0.4431	1.0	3.0
H_∞	1.8928	1.8958	0.0559	0.4784	1.0	3.0

The estimation of the correlation dimension using different kernel functions can vary. The estimation of the dimension for the Sierpinski carpet for different kernel functions is presented in Table 4. The table shows the theoretical dimension D_2 based on the parameters and its estimate $\widehat{D_2}$ together with the standard deviation sd. The recommended Ω range for L_4 regression is also included, where $f_{\min} = \log_{10} \Omega_{\min}$ and $f_{\max} = \log_{10} \Omega_{\max}$.

With the increasing dimension of the kernel function H_n, the standard deviation decreased. The estimates were unbiased in all cases; however, the most accurate estimation occurred for H_∞. The recommended intervals for the regression were the same in all cases, which means that they were independent of the kernel function; nevertheless, as will be seen later, they will be dependent on the theoretical dimension of the fractal structure.

All of the subsequent numerical experiments were performed in MATLAB by means of Monte Carlo simulation for $M = 10^5$ point pairs. At first, we tested the methodology for correlation dimension estimation for an n-dimensional Cantor dust with contraction coefficient a. This methodology provided an unbiased estimation of the correlation dimension in the whole range of possible theoretical dimensions for different contraction coefficients a. Almost the same behaviour was exhibited in the estimation of the corre-

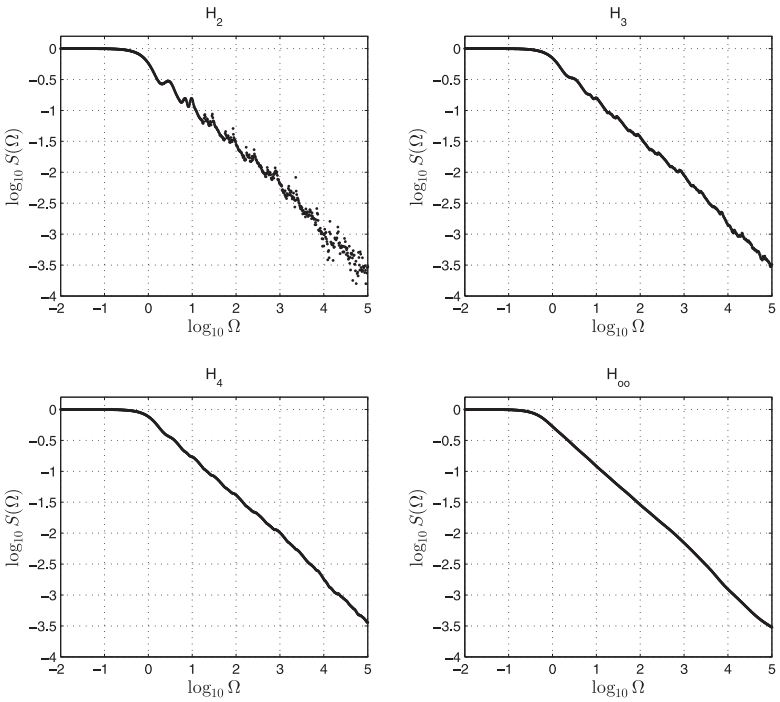


Fig. 3. Rotational spectra of the Sierpinski carpet.

Table 5
Analysis of the fractal sets with known dimension.

System	Parameters	D_2	$\widehat{D_2}$	sd	f_{min}	f_{max}
Cantor set	$a = 1/5,\ n = 1$	0.4307	0.4298	0.0396	0.0	2.0
Cantor set	$a = 1/3,\ n = 3$	1.8929	1.9065	0.0148	0.0	2.0
Levy flight	$\alpha = 1,\ n = 2$	1.0000	0.9718	0.0627	−3.0	0.0
Levy flight	$\alpha = 0.7,\ n = 3$	0.7000	0.6817	0.0323	−4.0	−2.0
fBm trajectory	$H = 0.75,\ n = 3$	1.3333	1.3334	0.0302	1.5	2.0
fBm trajectory	$H = 0.5, n = 4$	2.0000	1.9958	0.0203	1.3	1.8
fBm graph	$H = 0.3, n = 1$	1.7000	1.7003	0.0136	0.7	1.5
fBm graph	$H = 0.8, n = 1$	1.2000	1.1677	0.0512	0.8	2.0
Takagi function	$a = 0.75, n = 1$	1.5850	1.5889	0.0169	1.0	1.9
Takagi function	$a = 0.90, n = 1$	1.8480	1.8475	0.0149	1.1	1.8
Logistic map	[25]	0.5000	0.4948	0.0360	0.8	1.4
Rossler oscillator	[26]	2.0100	1.9845	0.0788	0.2	0.9
Lorenz system	[27]	1.2409	1.2388	0.0342	1.2	1.9

lation dimension of the trajectory of an n-dimensional Levy flight with the parameter $\alpha \in (0; 1)$. The lower and the upper boundary for the linear segment were shifted to the left when the theoretical dimension was lower. The graph of the fBm graph was also investigated to determine the efficiency of the D_2 estimator and had a theoretical dimension of $D_2 = 2 - H$. As a representative of the deterministic fractal functions, points from the Takagi function graph were generated. The Takagi function was dependent on the parameter $a \in (1/2; 1)$, and the theoretical correlation dimension was equal to $D_2 = 2 + \log_2 a$. The results from the estimation are shown in Table 5. The representatives of classical dynamic systems were also examined, namely, a logistic map [25] for $r = 3.56995$ with an estimated value of $D_2 = 0.500 \pm 0.005$; a Rossler oscillator [26] for $a = 0.1$, $b = 0.1$ and $c = 14$, with an estimated value of $D_2 = 2.01 \pm 0.01$; and a Lorenz-like system attractor [27] for $\alpha = 51/64$, $\beta = 396/256$ and $\theta = 109/64$, with $D_2 \in [1.2406, 1.2412]$.

8. Conclusion

The asymptotic behaviour of a rotational spectrum was investigated under the assumption of D_2 existence. Rotation in an infinite-dimensional space is recommended for correlation dimension estimation that is based on Monte Carlo simulation. As stated previously, there is a significant difference between traditional correlation integral behaviour and rotational spectrum, which can be seen on the basis of the log-log plot. The effect of spectrum stabilisation for $n \to \infty$ is also useful for D_2 estimation from relatively small uniform samples. However, the proposed method has a disadvantage in the experimental choice of the frequency range for regression, as in the case of the traditional approach.

Acknowledgement

The paper was written with the support of the Czech Technical University in Prague (grant no. SGS14/208/OHK4/3T/14).

References

[1] Edgar G. Measure, topology, and fractal geometry (undergraduate texts in mathematics). Springer; 2013. ISBN 03-877-4748-6.
[2] Falconer KJ. Fractal geometry : mathematical foundations and applications. Chichester: Wiley; 2003. ISBN 04-708-4862-6.
[3] Addison P. Fractals and chaos : an illustrated course. Bristol, UK Philadelphia, Pa: Institute of Physics Pub; 1997. ISBN 9780750304009.
[4] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D 1983;9(1–2):189–208. doi:10.1016/0167-2789(83)90298-1.
[5] Grassberger P, Procaccia I. Characterization of strange attractors. Phys Rev Lett 1983;50(5):346–9. doi:10.1103/physrevlett.50.346.
[6] Shayegh F, Sadri S, Amirfattahi R, Ansari-Asl K. A model-based method for computation of correlation dimension, lyapunov exponents and synchronization from depth-EEG signals. Comput Methods Programs Biomed 2014;113(1):323–37. doi:10.1016/j.cmpb.2013.08.014.
[7] Rawal K, Saini BS, Saini I. Adaptive correlation dimension method for analysing heart rate variability during the menstrual cycle. Australas Phys Eng Sci Med 2015;38(3):509–23. doi:10.1007/s13246-015-0369-y.
[8] Kalauzi A, Vuckovic A, Bojić T. Topographic distribution of EEG alpha attractor correlation dimension values in wake and drowsy states in humans. Int J Psychophysiol 2015;95(3):278–91. doi:10.1016/j.ijpsycho.2014.11.008.
[9] Liu Y, Yu Z, Zeng M, Wang S. Dimension estimation using weighted correlation dimension method. Discrete Dyn Nat Soc 2015;2015:1–10. doi:10.1155/2015/837185.
[10] Michalak KP. How to estimate the correlation dimension of high-dimensional signals? Chaos 2014;24(3):033118. doi:10.1063/1.4891185.
[11] Hongying Y, Duanfeng J. Mathematical modelling: v. 2: proceedings of first international conference on modelling and simulation. World Academic Union Ltd; 2008. ISBN 1846260426.
[12] Grafakos L. Classical fourier analysis: graduate texts in mathematics. Springer; 2014. ISBN 1493911937.
[13] Talebinejad M, Chan AD, Miri A, Dansereau RM. Fractal analysis of surface electromyography signals: a novel power spectrum-based method. J Electromyography Kinesiol 2009;19(5):840–50. doi:10.1016/j.jelekin.2008.05.004.
[14] Wen H, Liu Z. Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. Brain Topogr 2015;29(1):13–26. doi:10.1007/s10548-015-0448-0.
[15] Churnside JH, Wilson JJ. Power spectrum and fractal dimension of laser backscattering from the ocean. J Opt Soc Am A 2006;23(11):2829. doi:10.1364/josaa.23.002829.
[16] Debye P. Zerstreuung von röntgenstrahlen. Ann Phys 1915;351(6):809–23. doi:10.1002/andp.19153510606.
[17] Krantz SG. Handbook of complex variables. Birkhäuser; 1999. ISBN 0817640118.
[18] Parzen E. On estimation of a probability density function and mode. Ann Mat Stat 1962;33(3):1065–76. doi:10.1214/aoms/1177704472.
[19] Fischer R. The adaptive resolution concept in form-free distribution estimation. In: Kluge W, editor. Proceedings of the Workshop on Physics and Computer Science. Heidelberg; 1999.
[20] von der Linden VDW, Fischer R. Adaptive kernels and Occam's razor in inversion problems. In: Sears M, Nedeljkovic V, Pendock N, Sibisi S, editors. Proceedings of the Maximum Entropy Conference. NBM printers, Port Elizabeth, South Africa; 1997.
[21] Olson EJ, Robinson JC, Sharples N. Generalised cantor sets and the dimension of products. Math Proc Cambridge Philos Soc 2015;160(01):51–75. doi:10.1017/s0305004115000584.
[22] Hughes BD. Random walks and random environments: volume 1: random walks. Clarendon Press; 1995. ISBN 0198537883.
[23] Wu DS, Xiao YM. Dimensional properties of fractional brownian motion. Acta Math Sin Engl Ser 2007;23(4):613–22. doi:10.1007/s10114-005-0928-3.
[24] Takeo F. Box-counting dimension of graphs of generalized takagi series. Jpn J Ind Appl Math 1996;13(2):187–94. doi:10.1007/bf03167241.
[25] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D 1983;9(1–2):189–208.
[26] Peitgen H-O, Jürgens H, Saupe D. Chaos and fractals: new frontiers of science. Springer; 1993. ISBN 0387979034.
[27] Galatolo S, Nisoli I. Rigorous computation of invariant measures and fractal dimension for maps with contracting fibers: 2d lorenz-like maps. Ergodic Theor Dynam Syst 2015;36(06):1865–91. doi:10.1017/etds.2014.145.

Chaos, Solitons and Fractals 114 (2018) 536–541

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Frontiers

Translation and rotation invariant method of Renyi dimension estimation

Martin Dlask*, Jaromir Kukal

Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Department of Software Engineering, Trojanova 13, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 11 November 2017
Revised 18 July 2018
Accepted 24 July 2018

Keywords:
Parzen estimate
Renyi entropy
Monte carlo
Renyi dimension

ABSTRACT

A fractal dimension is a non-integer characteristic that measures the space filling of an arbitrary set. The conventional methods usually provide a biased estimation of the fractal dimension, and therefore it is necessary to develop more complex methods for its estimation. A new characteristic based on the Parzen estimate formula is presented, and for the analysis of correlation dimension, a novel approach that employs the log-linear dependence of a modified Renyi entropy is used. The new formula for the Renyi entropy has been investigated both theoretically and experimentally on selected fractal sets.

1. Introduction

A fractal dimension is a non-integer characteristic that allows to measure the space-filling property of any set in Euclidean space. There are several definitions of dimension. The most general definition is the Hausdorff dimension [1] based on the Hausdorff measure and covering of the set with smaller sets with different radii. The similarity dimension [2] is based on the scaling property of fractal sets and is used only for the analysis of simple self-similar sets. The parameterized type of a fractal dimension is the Renyi dimension [1,3] that is the main object of interest in this paper.

The calculation of the Renyi dimension is based on the Renyi entropy [4] H_α, which is a generalisation of the Shannon (H_1), Hartley (H_0) and collision (H_2) entropies. The α-entropy is defined for $\alpha \geq 0$ as follows

$$H_\alpha(\vec{p}) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{k} p_i^\alpha \tag{1}$$

for $\alpha \neq 1$ and

$$H_1 = \lim_{\alpha \to 1} H_\alpha = -\sum_{i=1}^{k} p_i \ln p_i \tag{2}$$

where k is the number of events and p_i are their probabilities satisfying $\sum_{i=1}^{k} p_i = 1$. The formulas (2) and (1) are frequently used in most sources, but they describe only a finite set of events with a possible extension to a countable case. A more general form which includes also an uncountable case is defined as

$$H_\alpha(\vec{p}) = \frac{1}{1-\alpha} \ln \mathrm{E}\left(p^{\alpha-1}\right) \tag{3}$$

and

$$H_1 = \mathrm{E}\left(-\ln p\right). \tag{4}$$

Based on the definition of α-entropy, the Renyi dimension is defined as

$$D_\alpha = \lim_{\epsilon \to 0^+} \frac{H_\alpha}{-\ln \epsilon} \tag{5}$$

where D_0, D_1 and D_2 are called the capacity, information and correlation dimension, respectively. In this case, ϵ is a scaling parameter that influences the probabilities p_i.

The methods that are used to estimate the Renyi dimension are usually different for different parameters α. The capacity dimension for $\alpha = 0$ is usually estimated via the box-counting method [5] or the Minkowski covering method [1]. The particular type of dimension for $\alpha = 2$ is called the correlation dimension and was introduced first in [6]. There are several methods to estimate the correlation dimension including the traditional approach in [7] or the spectral approach in [8].

The determination of the Renyi dimension is based on Renyi entropy estimates, which is biased in general. The second and more general problem is how to sample the point set. Our approach is focused only on the Lebesgue measure sets where uniform sampling is defined. When these conditions are not guaranteed, such as when the geometric structure of the set is inhomogeneous, we

* Corresponding author.
E-mail address: martindlask@centrum.cz (M. Dlask).

https://doi.org/10.1016/j.chaos.2018.07.030
0960-0779/

Fig. 1. Density of Parzen estimate and corresponding $\mathcal{S}_M$.

can only test the hypothesis of unbiasedness for the given theoretical value D_α i.e., $H_0 : \widehat{D_\alpha} = D_\alpha$. In such cases, it was shown that the correlation sum represents an unbiased estimator of D_2 with respect to approach in [9,10].

Recently, there were efforts to improve the estimation of the capacity dimension of binary images [11,12] and to estimate this dimension of the set of possible singular points in the space-time of suitable weak solutions to the Navier–Stokes equations [13,14]. The correlation dimension is widely used in biomedicine for electroencephalography signal analysis [15,16] or in cardiology [17]. Economical data are also often the subject of the correlation dimension analysis, for example financial markets [18], and especially capital markets [19].

2. Parzen estimate with ball kernel

This section utilizes the Parzen estimate for the derivation of the density function of elements of the Lebesgue measurable set $\mathcal{F} \subset \mathbb{R}^n$. Supposing the existence of n-dimensional distribution function ϕ of points $\vec{x} \in \mathcal{F}$ i.e., $\vec{x} \sim \phi$, it is possible to define a sample of points

$$\Phi = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_M}\} \subset \mathcal{F} \tag{6}$$

that are uniformly generated from $\mathcal{F}$, i.e., $\vec{x_k} \sim \mathrm{U}(\mathcal{F}) \equiv \phi$. For any point $\vec{x} \in \mathbb{R}^n$, we define its ϵ-neighbourhood, i.e., a ball with radius ϵ as

$$\mathcal{B}(\vec{x}, \epsilon) = \{\vec{y} \in \mathbb{R}^n : \|\vec{y} - \vec{x}\|_2 \le \epsilon\} \tag{7}$$

for any $\epsilon > 0$. The volume V_* of the ball can be expressed as

$$V_* = V_n \cdot \epsilon^n \tag{8}$$

where V_n is the volume of an n-dimensional unit ball. The density estimate will be based on the elementary distribution

$$f_0(\vec{x}, \epsilon) = \frac{\mathrm{I}(\|\vec{x}\|_2 \le \epsilon)}{V_*} \tag{9}$$

using the indicator function $\mathrm{I}(\ldots)$. We can use Parzen's [20] formula

$$f(\vec{x}, \Phi, \epsilon) = \frac{1}{M} \sum_{k=1}^{M} f_0(\vec{x} - \vec{x_k}, \epsilon) \tag{10}$$

to obtain a consistent estimate of ϕ. However, we will apply (10) to a discontinuous distribution on $\mathcal{F}$ to obtain new formulas for the Renyi dimension estimation.

The probability density estimate (10) is visualised on Fig. 1 in grayscale. The white area represents the regions where this function equals zero and the darker areas depict the intersection of several balls centred at points from the set sample Φ. The balls can be also used for the traditional definition of the Minkowski sausage [2] as

$$\mathcal{S} = \bigcup_{\vec{x} \in \mathcal{F}} \mathcal{B}(\vec{x}, \epsilon). \tag{11}$$

The sample set Φ is useful for its finite approximation

$$\mathcal{S} \approx \mathcal{S}_M = \bigcup_{k=1}^{M} \mathcal{B}(\vec{x_k}, \epsilon). \tag{12}$$

3. Renyi entropy estimate

Our novel estimate of the Renyi entropy is based on the differential entropy

$$H_\alpha = \frac{1}{1-\alpha} \ln \int_{\vec{x} \in \mathbb{R}^n} f^\alpha(\vec{x}) \mathrm{d}\vec{x} \tag{13}$$

for $\alpha \in \mathbb{R}_0^+ \setminus \{1\}$ and the Parzen estimate $f(\vec{x})$ that is scaled by $\epsilon > 0$. To avoid negative entropy values, we define the modified Renyi entropy as

$$H_\alpha^*(\Phi, \epsilon) = \frac{\ln J(\Phi, \alpha, \epsilon) - \ln J_0(\alpha, \epsilon)}{1-\alpha}. \tag{14}$$

for $\alpha \ge 0$ and $\alpha \ne 1$ where

$$J(\Phi, \alpha, \epsilon) = \int_{\vec{x} \in \mathbb{R}^n} f^\alpha(\vec{x}, \Phi, \epsilon) \mathrm{d}\vec{x} \tag{15}$$

and

$$J_0(\alpha, \epsilon) = \int_{\vec{x} \in \mathbb{R}^n} f_0^\alpha(\vec{x}, \epsilon) \mathrm{d}\vec{x} = V_*^{1-\alpha}. \tag{16}$$

Using the expected value of $v(\vec{x})$ for $\vec{x} \sim g$ as

$$\mathop{\mathrm{E}}_{\vec{x} \sim g} v(\vec{x}) = \int_{\vec{x} \in \mathbb{R}^n} v(\vec{x}) g(\vec{x}) \mathrm{d}\vec{x}, \tag{17}$$

the first term can be simplified as

$$\begin{aligned} J = J(\Phi, \alpha, \epsilon) &= \int_{\vec{x} \in \mathbb{R}^n} f^{\alpha-1}(\vec{x}, \Phi, \epsilon) \cdot f(\vec{x}, \Phi, \epsilon) \mathrm{d}\vec{x} \\ &= \mathop{\mathrm{E}}_{\vec{x} \sim f_\Phi} f^{\alpha-1}(\vec{x}, \Phi, \epsilon) \end{aligned} \tag{18}$$

We define the degeneracy of $\vec{x} \in \mathbb{R}^n$ as

$$G(\vec{x}, \Phi, \epsilon) = \sum_{k=1}^{M} \mathrm{I}(\|\vec{x} - \vec{x_k}\|_2 \le \epsilon) \tag{19}$$

holding that $G(\vec{x}, \Phi, \epsilon) \in \{0, \ldots, M\}$. Recall that the probability density function $f(\vec{x})$ is

$$f(\vec{x}, \Phi, \epsilon) = \frac{1}{M \cdot V_*} \sum_{k=1}^{M} \mathrm{I}(\|\vec{x} - \vec{x_k}\|_2 \le \epsilon) = \frac{G(\vec{x}, \Phi, \epsilon)}{M \cdot V_*}. \tag{20}$$

Therefore

$$J = \mathop{\mathrm{E}}_{\vec{x} \sim f_\Phi} \left(\frac{G(\vec{x}, \Phi, \epsilon)}{M \cdot V_*} \right)^{\alpha-1} = M^{1-\alpha} V_*^{1-\alpha} \mathop{\mathrm{E}}_{\vec{x} \sim f_\Phi} G^{\alpha-1}(\vec{x}, \Phi, \epsilon) \tag{21}$$

and subsequently also the modified Renyi entropy is

$$H_\alpha^*(\Phi, \epsilon) = \frac{\ln J - \ln J_0}{1-\alpha} \tag{22}$$

$$= \frac{(1-\alpha)\ln M + (1-\alpha)\ln V_* + \ln \mathrm{E}\, G^{\alpha-1}(\vec{x}, \Phi, \epsilon) - (1-\alpha)\ln V_*}{1-\alpha} \tag{23}$$

The resulting modified entropy equals

$$H_\alpha^*(\Phi, \epsilon) = \ln M + \frac{\ln \mathrm{E}\, G^{\alpha-1}(\vec{x}, \Phi, \epsilon)}{1-\alpha} \tag{24}$$

for $\alpha > 0$ and $\alpha \ne 1$.

4. Basic properties and particular cases

In this and the following sections, the degeneracy of $\vec{x} \in \mathbb{R}^n$ will be denoted as $G(\vec{x})$ instead of $G(\vec{x}, \Phi, \epsilon)$. When $\vec{x} \in \mathbb{R}^n$, the degeneracy $G(\vec{x}) \in \{0, \dots, M\}$, but for $\vec{x} \in \mathcal{S}_M$, the degeneracy fulfils $G(\vec{x}) \in \{1, \dots, M\}$. The modified Renyi entropy follows $0 \le H_\alpha^* \le \ln M$. This entropy is a translational and rotational invariant, as it is easy to prove from (19) and (24). For the particular cases of α, one can derive the

- Modified Hartley entropy for $\alpha = 0$ as

$$H_0^* = \ln M + \ln \mathrm{E} G^{-1}(\vec{x}), \tag{25}$$

- Modified Shannon entropy as a limit for $\alpha \to 1$ i.e.

$$H_1^* = \lim_{\alpha \to 1} H_\alpha^* = \ln M - \mathrm{E}\ \ln G(\vec{x}), \tag{26}$$

- Modified collision entropy for $\alpha = 2$ as

$$H_2^* = \ln M - \ln \mathrm{E} G(\vec{x}), \tag{27}$$

- Modified minimum entropy as a limit for $\alpha \to \infty$ as

$$H_\infty^* = \lim_{\alpha \to +\infty} H_\alpha^* = \ln M - \ln \max G(\vec{x}). \tag{28}$$

where the expected values are over $\vec{x} \sim f$. If the derivative $\frac{\partial H^*}{\partial \alpha}$ exists, it is always less or equal to zero, as it is easy to prove.

Moreover, the modified Renyi entropy H_α^* can be used for an alternative definition of the dimension as

$$D_\alpha^* = \lim_{\epsilon \to 0^+} \frac{H_\alpha^*(\epsilon)}{-\ln \epsilon} \tag{29}$$

for a given $\mathcal{F}$ as an analogy to formula (5).

5. Monte carlo approach

Basic properties of the finite sample Φ have been collected in the previous sections. Their direct application to the given data set is useful for the estimation of $H_\alpha^*(\epsilon)$ and consequently for the D_α^* estimation. Using the operator U of uniform sampling, the approximation of the Renyi entropy can be achieved via a Monte Carlo technique in the following way:

1. At first, the sample index is generated uniformly $k \sim \mathrm{U}(\{1, \dots, M\})$.
2. The point $\vec{x}$ is generated uniformly from the ϵ-ball centred at $\vec{x_k}$ as $\vec{x} \sim \mathrm{U}(\mathcal{B}(\vec{x_k}, \epsilon))$.
3. The subsequent degeneration is calculated using (19).

The entropy H_α^* is calculated as an average of the degenerations using (24), (26) or (28) depending on α. The first two steps generate $\vec{x} \sim f$, of course. Assuming the entropy estimate H_α^* fulfils $H_\alpha^* \propto \epsilon^{-D_\alpha^*}$ for small $\epsilon > 0$, we can use it for the estimation of D_α^* using the model

$$H_\alpha^*(\epsilon) = A - D_\alpha^* \ln \epsilon \tag{30}$$

for small ϵ and satisfying linear dependency H_α^* on $\ln \epsilon$. The aim of this study is to demonstrate that D_α^* is an unbiased estimate of D_α for large M.

6. Relationship to capacity and correlation dimension

The capacity (D_0) and correlation (D_2) dimensions are defined for any Lebesgue measurable set $\mathcal{F}$. The only possibility how to compare D_α^* with D_α is to come back from the sample Φ to the original set $\mathcal{F}$. The sample Φ is a finite set with $D_\mathrm{H} = D_0 = D_2 = 0$, of course. We will study the particular cases of D_α^* for $\alpha = 0$ and $\alpha = 2$ in the case of the measurable set $\mathcal{F}$.

6.1. Relationship to D_0

The Renyi dimension is the characteristic that has an important relationship to the Minkowski-Bouligard capacity dimension. The capacity dimension can be defined [1] based on the Minkowski sausage as

$$D_0 = n - \lim_{\epsilon \to 0^+} \frac{\ln \mathrm{vol}(\mathcal{S})}{\ln \epsilon} \tag{31}$$

where $\mathcal{S}$ is defined in (11) and $\mathrm{vol}(\mathcal{S}) = \int_{\vec{x} \in \mathcal{S}}$ is its volume. Supposing the existence of D_0, we can directly calculate

$$J_0 = \int_{\mathbb{R}^n} f_0^0(\vec{x}) \mathrm{d}\vec{x} = V_* = V_n \cdot \epsilon^n, \tag{32}$$

and also the density

$$f(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}) \tag{33}$$

and

$$f^0(\vec{x}) = \mathrm{I}\left(\mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}) > 0 \right) = \mathrm{I}\left(\bigvee_{\vec{y} \in \mathcal{F}} \|\vec{x} - \vec{y}\|_2 \le \epsilon \right) = \mathrm{I}(\vec{x} \in \mathcal{S}). \tag{34}$$

Therefore, the function J can be expressed as

$$J = \int_{\mathbb{R}^n} f^0(\vec{x}) \mathrm{d}\vec{x} = \mathrm{vol}(\mathcal{S}). \tag{35}$$

The resulting modified Hartley entropy equals

$$\begin{aligned} H_0^*(\epsilon) &= \ln \frac{\int_{\mathbb{R}^n} f^0(\vec{x}) \mathrm{d}\vec{x}}{\int_{\mathbb{R}^n} f_0^0(\vec{x}) \mathrm{d}\vec{x}} = \ln \mathrm{vol}(\mathcal{S}) - \ln V_* \\ &= \ln \mathrm{vol}(\mathcal{S}) - \ln V_n - n \ln \epsilon. \end{aligned} \tag{36}$$

Now, it is clear that

$$\begin{aligned} D_0^* &= \lim_{\epsilon \to 0^+} \frac{H_0^*(\epsilon)}{-\ln \epsilon} = \lim_{\epsilon \to 0^+} \frac{\ln \mathrm{vol}(\mathcal{S}) - \ln V_n - n \ln \epsilon}{-\ln \epsilon} \\ &= n - \lim_{\epsilon \to 0^+} \frac{\ln \mathrm{vol}(\mathcal{S})}{\ln \epsilon} = D_0. \end{aligned} \tag{37}$$

Therefore, D_0^* obtained from the modified Hartley entropy $H_0^*(\epsilon)$ is equivalent to the capacity dimension D_0 of the measurable set $\mathcal{F}$.

6.2. Relationship to D_2

The correlation dimension of $\mathcal{F}$ is defined as

$$D_2 = \lim_{\epsilon \to 0^+} \frac{\ln C(\epsilon)}{\ln \epsilon} \tag{38}$$

where

$$C(\epsilon) = \mathop{\mathrm{E}}_{\vec{y}, \vec{z} \sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y} - \vec{z}\|_2 \le \epsilon) \tag{39}$$

is the correlation integral. Supposing the existence of D_2, recall that

$$f_0(\vec{x}) = \frac{\mathrm{I}(\|\vec{x}\|_2 \le \epsilon)}{V_n \epsilon^n} \tag{40}$$

and by means of integrating the elementary distribution over the space we get

$$J_0 = \int_{\mathbb{R}^n} f_0^2(\vec{x}) \mathrm{d}\vec{x} = \frac{1}{V_n^2 \epsilon^{2n}} V_n \epsilon^n = V_n^{-1} \epsilon^{-n}. \tag{41}$$

In the finite case, we have

$$f(\vec{x}) = \frac{1}{m} \sum_{k=1}^{m} f_0(\vec{x} - \vec{x_k}), \tag{42}$$

which can be generalized to

$$f(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y} \sim \mathrm{U}(\mathcal{F})} f_0(\vec{x} - \vec{y}). \tag{43}$$

Therefore,

$$f^2(\vec{x}) = \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} f_0(\vec{x}-\vec{y})\cdot f_0(\vec{x}-\vec{z}) \tag{44}$$

and

$$J = \int_{\mathbb{R}^n} f^2(\vec{x})\mathrm{d}\vec{x} = \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} Z(\vec{y},\vec{z}) \tag{45}$$

where

$$Z(\vec{y},\vec{z}) = \int_{\mathbb{R}^n} f_0(\vec{x}-\vec{y})\cdot f_0(\vec{x}-\vec{z})\mathrm{d}\vec{x}. \tag{46}$$

Denoting the distance $d = \|\vec{y}-\vec{z}\|_2$, we can evaluate

$$\begin{aligned} Z(\vec{y},\vec{z}) &= V_n^{-2}\epsilon^{-2n}\int_{\mathbb{R}^n} \mathrm{I}(\|\vec{x}-\vec{y}\|_2 \le \epsilon)\mathrm{I}(\|\vec{x}-\vec{z}\|_2 \le \epsilon)\mathrm{d}\vec{x} \\ &= V_n^{-2}\epsilon^{-2n}W_n(d,\epsilon) \end{aligned} \tag{47}$$

where $W_n(d, \epsilon)$ is the volume of two hyper-ball intersections in the case of the mutual center distance d and radii ϵ. Using n-dimensional analytical geometry, we obtain

$$\begin{aligned} W_n(d,\epsilon) &= 2\int_{d/2}^{\epsilon} V_{n-1}\left(\epsilon^2 - r^2\right)^{\frac{n-1}{2}}\mathrm{d}r \\ &= 2V_{n-1}\epsilon^n\int_{d/2\epsilon}^{1}\left(1-r^2\right)^{\frac{n-1}{2}}\mathrm{d}r \end{aligned} \tag{48}$$

and after substitution $r = \cos\phi$, we get

$$W_n(d,\epsilon) = 2V_{n-1}\epsilon^n\int_0^{\arccos(d/2\epsilon)} \sin^n\phi \ \ \mathrm{d}\phi. \tag{49}$$

Moreover,

$$W_n(0,\epsilon) = 2V_{n-1}\epsilon^n\int_0^{\pi/2} \sin^n\phi \ \ \mathrm{d}\phi = V_n\cdot\epsilon^n \tag{50}$$

which is also the volume of the n-dimensional ball of radius ϵ. Therefore, we can express the Z function as

$$Z(\vec{y},\vec{z}) = V_n^{-1}\epsilon^{-n}\frac{\int_0^{\arccos(d/2\epsilon)}\sin^n\phi \ \ \mathrm{d}\phi}{\int_0^{\pi/2}\sin^n\phi \ \ \mathrm{d}\phi} \tag{51}$$

and the entropy is

$$H_2^*(\epsilon) = -\ln \mathop{\mathrm{E}}_{\vec{x},\vec{y}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{x}-\vec{y}\|_2,\epsilon) \tag{52}$$

where

$$S_n(d,\epsilon) = \frac{\int_0^{\arccos(d/2\epsilon)}\sin^n\phi \ \ \mathrm{d}\phi}{\int_0^{\pi/2}\sin^n(\phi)\mathrm{d}\phi} \tag{53}$$

for $0 \le d < 2\epsilon$ and $S_n(d,\epsilon) = 0$ for $d \ge 2\epsilon$. Let

$$Q_n = \frac{\int_0^{\pi/3}\sin^n\phi \ \ \mathrm{d}\phi}{\int_0^{\pi/2}\sin^n\phi \ \ \mathrm{d}\phi} \in (0,1) \tag{54}$$

be the value of $S_n(\epsilon, \epsilon)$. When $0 \le d \le \epsilon$, we can estimate the ratio as

$$Q_n \le S_n(d,\epsilon) \le 1. \tag{55}$$

For $\epsilon < d \le 2\epsilon$, we have $0 \le S_n(d, \epsilon) < Q_n$. Therefore, we can underestimate

$$S_n(d,\epsilon) \ge \mathrm{I}(d \le \epsilon)\cdot Q_n \tag{56}$$

and an adequate upper estimate is

$$S_n(d,\epsilon) \le \mathrm{I}(d \le \epsilon) + (\mathrm{I}(d \le 2\epsilon) - \mathrm{I}(d \le \epsilon))\cdot Q_n \tag{57}$$

$$\begin{aligned} &(1-Q_n)\cdot\mathrm{I}(d \le \epsilon) + Q_n\cdot\mathrm{I}(d \le 2\epsilon) \le (1-Q_n)\cdot\mathrm{I}(d \le 2\epsilon) \\ &+Q_n\cdot\mathrm{I}(d \le 2\epsilon) = \mathrm{I}(d \le 2\epsilon). \end{aligned} \tag{58}$$

We can continue in the estimation to obtain

$$\begin{aligned} &\mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y}-\vec{z}\|_2 \le \epsilon)\cdot Q_n \le \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{y}-\vec{z}\|_2,\epsilon) \\ &\le \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} \mathrm{I}(\|\vec{y}-\vec{z}\|_2 \le 2\epsilon) \end{aligned} \tag{59}$$

and therefore

$$L(\epsilon) \le \mathop{\mathrm{E}}_{\vec{y},\vec{z}\sim \mathrm{U}(\mathcal{F})} S_n(\|\vec{y}-\vec{z}\|_2,\epsilon) \le U(\epsilon) \tag{60}$$

where the lower bound equals

$$L(\epsilon) = Q_n\cdot C(\epsilon) \tag{61}$$

and the appropriate upper bound equals

$$U(\epsilon) = C(2\epsilon). \tag{62}$$

For all $0 < \epsilon < 1$, the following inequalities hold

$$\frac{\ln U(\epsilon)}{\ln\epsilon} \le \frac{H_2^*(\epsilon)}{-\ln\epsilon} \le \frac{\ln L(\epsilon)}{\ln\epsilon}. \tag{63}$$

We can calculate

$$\lim_{\epsilon\to 0^+}\frac{\ln L(\epsilon)}{\ln\epsilon} = \lim_{\epsilon\to 0^+}\left(\frac{\ln Q_n}{\ln\epsilon} + \frac{\ln C(\epsilon)}{\ln\epsilon}\right) = D_2 \tag{64}$$

and also for the upper bound

$$\begin{aligned} \lim_{\epsilon\to 0^+}\frac{\ln U(\epsilon)}{\ln\epsilon} &= \lim_{\epsilon\to 0^+}\frac{\ln C(2\epsilon)}{\ln\epsilon} = \lim_{\epsilon\to 0^+}\frac{\ln C(\epsilon)}{\ln\frac{\epsilon}{2}} \\ &= \lim_{\epsilon\to 0^+}\frac{\ln C(\epsilon)}{\ln\epsilon}\cdot\lim_{\epsilon\to 0^+}\frac{\ln\epsilon}{\ln\epsilon - \ln 2} = D_2. \end{aligned} \tag{65}$$

Therefore,

$$D_2^* = \lim_{\epsilon\to 0^+}\frac{H_2^*(\epsilon)}{-\ln\epsilon} = D_2. \tag{66}$$

As a conclusion, when D_0, D_2 exist for a given set $\mathcal{F}$, the equalities $D_0^* = D_0$, $D_2^* = D_2$ have been proven.

7. Experimental part

Computer experiments can be realized only on the finite sample Φ with three aims:

- verify hypothesis $H_0 : \widehat{D}_0^* = D_0$ experimentally,
- verify hypothesis $H_0 : \widehat{D}_2^* = D_2$ experimentally,
- evaluate $\widehat{D}_\alpha^*$ in other cases where D_α is known theoretically or is referenced.

The estimation of the Renyi dimension D_α^* will be performed for $\alpha \in [0; 2]$. Supposing the model (30) with additional Gaussian noise $e \sim \mathcal{N}(0;\sigma^2)$ in the form

$$H_\alpha^* = A - D_\alpha^*\ln\epsilon + e \tag{67}$$

we can use the least squares method for the D_α^* estimation using different values ϵ_i for $i = 1,\ldots,N$. We suggest to use a geometrically increasing series of ϵ_i generated by the formula

$$\epsilon_i = 10^{f_{\min}+(i-1)\Delta f} \tag{68}$$

with $N = \lfloor (f_{\max} - f_{\min})/\Delta f \rfloor + 1$.

The novel algorithm was tested on sets with known capacity dimensions. Four traditional deterministic fractal sets were studied using recursive random point generation of depth 100:

- Cantor set [1] with the contraction parameter $0 < a < 1/2$ and $n = 1$ with the Hausdorff dimension

$$D_{\mathrm{H}} = -\frac{\ln 2}{\ln a} \tag{69}$$

Table 1
Capacity dimension estimation.

System	a	D_0	$\widehat{D_0^*}$	sd	p-value	$f_{\min}$	$f_{\max}$
Cantor set	1/4	0.5000	0.5059	0.0070	0.3993	−3.0	−1.0
Cantor set	1/3	0.6309	0.6327	0.0034	0.5965	−3.0	−1.0
Cantor dust	1/4	1.0000	0.9834	0.0157	0.2937	−2.0	0.0
Cantor dust	1/3	1.2619	1.2547	0.0133	0.5883	−2.0	0.0
Even numbers set	–	0.6990	0.7030	0.0148	0.7870	−4.0	−1.0
Sierpinki carpet	1/3	1.8928	1.8894	0.0059	0.2843	−2.0	−1.0
Sierpinki carpet	1/4	1.5000	1.4901	0.0148	0.2514	−2.0	−1.0

Table 2
Correlation dimension estimation.

System	a	D_2	$\widehat{D_2^*}$	sd	p-value	$f_{\min}$	$f_{\max}$
Cantor set	1/4	0.500	0.4974	0.0034	0.2236	−3.0	−1.0
Cantor set	1/3	0.6309	0.6286	0.0047	0.3124	−3.0	−1.0
Cantor dust	1/4	1.0000	0.9863	0.0221	0.2676	−2.0	−1.0
Cantor dust	1/3	1.2619	1.2630	0.0269	0.4840	−2.0	0.0
Even numbers set	–	0.6990	0.6991	0.0038	0.4896	−4.0	−1.0
Sierpinki carpet	1/3	1.8928	1.8964	0.0083	0.3325	−2.0	−1.0
Sierpinki carpet	1/4	1.5000	1.5053	0.0064	0.2032	−2.0	−1.0

- Cantor dust [21] with the contraction parameter $0 < a < 1/2$ and $n = 2$ with the Hausdorff dimension

$$D_{\mathrm{H}} = -2\frac{\ln 2}{\ln a} \tag{70}$$

- Even the digits set [1] that contains numbers from (0, 1) with even digits and the Hausdorff dimension

$$D_{\mathrm{H}} = -\frac{\log 5}{\log 10} \tag{71}$$

- Sierpinski carpet [1] with the contraction parameter $0 < a < 1/2$ and $n = 2$ with the Hausdorff dimension

$$D_{\mathrm{H}} = -\frac{\ln 8}{\ln a} \tag{72}$$

Since all the mentioned sets are self-similar and fulfil the open set condition [22], their Hausdorff dimension equals the Renyi dimension for any eligible parameter α, e.g. $D_{\mathrm{H}} = D_0 = D_2$. The results of the capacity dimension estimation are provided in Table 1 and the estimates in the case of the correlation dimension are in Table 2 for $\Delta f = 0.05$ and $M = 10^5$. The theoretical capacity (correlation) dimension is denoted D_0 (D_2), whereas its estimate is $\widehat{D_0^*}$ ($\widehat{D_2^*}$) together with its standard deviation sd. The range for the choice of $\ln\epsilon$ is recommended to be in the interval $[f_{\min}; f_{\max}]$.

A one-sample, two-sided t-test has been used to prove the unbiasedness of the dimension estimates level 0.05. As seen in Tabs. 1 and 2 the hypotheses $H_0 : \widehat{D_\alpha^*} = D_\alpha$ have been accepted in all cases.

The graph of De Wijs's fractal [23] with the parameter a is a kind of multifractal that has the Renyi dimension dependent on the dimension parameter α. The corresponding Renyi dimension equals

$$D_\alpha = \frac{1}{1-\alpha}\log_2\left(a^\alpha + (1-a)^\alpha\right) \tag{73}$$

for $0 < a < 1/2$ and $\alpha \in [0; 1) \cup (1, \infty)$ with the particular case

$$D_1 = \lim_{\alpha\to 1} D_\alpha = -a\log_2 a - (1-a)\log_2(1-a). \tag{74}$$

The D_α^* has been estimated for $\alpha \in \{0, 1/2, 1, 3/2, 2\}$ and the testing results are included in Table 3.

The one-sample, two-sided t-test has been also used to prove the unbiasedness of the De Wijs's fractal, the hypotheses $H_0 : \widehat{D_\alpha^*} = D_\alpha$ have been again accepted in all cases.

One of the traditional methods on how to estimate the capacity dimension D_0 is called box-counting [24]. It is based on counting points from the sample Φ using an n-dimensional rectangular grid of size $a > 0$. Using the grid, there are always k non-empty boxes consisting of $M_1, M_2, \ldots, M_k \in \mathbb{N}$ points satisfying $\sum_{j=1}^{k} M_j = M$. The basic form of box-counting calculates the Hartley entropy estimate according to (1) as $\widehat{H_0}(\alpha) = \ln k$ which is the logarithm of covering an element number. The box-counting estimate of D_0 is obtained from the model (67). It is also possible to estimate the general Renyi entropy D_α using the approximation $p_j = M_j/M$ and formulas (2) and (1).

Table 3
De Wijs's fractal dimensions.

α	a	D_α	$\widehat{D_\alpha^*}$	sd	p-value	$f_{\min}$	$f_{\max}$
0	$a = 1/3$	1.0000	0.9908	0.0058	0.1127	−6.0	−4.0
0	$a = 1/4$	1.0000	0.9890	0.0087	0.2062	−6.0	−4.0
0	$a = 1/6$	1.0000	0.9780	0.0143	0.1240	−6.0	−4.0
1/2	$a = 1/3$	0.9581	0.9574	0.0062	0.4550	−5.5	−3.5
1/2	$a = 1/4$	0.9000	0.8921	0.0103	0.2215	−5.5	−3.5
1/2	$a = 1/6$	0.8035	0.7895	0.0159	0.1893	−5.0	−3.0
1	$a = 1/3$	0.9183	0.9158	0.0060	0.6769	−4.0	−2.0
1	$a = 1/4$	0.8250	0.8259	0.0098	0.9269	−4.0	−2.0
1	$a = 1/6$	0.6500	0.6387	0.0217	0.6026	−3.0	−1.0
3/2	$a = 1/3$	0.8814	0.8749	0.0099	0.2557	−4.0	−2.0
3/2	$a = 1/4$	0.7376	0.7255	0.0153	0.2145	−4.0	−2.0
3/2	$a = 1/6$	0.5419	0.5234	0.0209	0.1880	−3.0	−1.0
2	$a = 1/3$	0.8480	0.8359	0.0189	0.5220	−3.0	−1.0
2	$a = 1/4$	0.6781	0.6687	0.0205	0.6466	−3.0	−1.0
2	$a = 1/6$	0.4695	0.4552	0.0235	0.5429	−2.0	0.0

Discrete dynamic systems with chaotic behaviour generate fractal trajectories and attractors with a nonlinear character. The investigation of this kind of sets can be performed in two ways – the first option is to investigate the dimension in the original state space, the second option is to use Whitney's theorem [25] and estimate it in a reconstructed space. Generally, the n-dimensional discrete dynamical process has an internal state $\vec{x}_j \in \mathbb{R}^n$ and output $y_j \in \mathbb{R}$ for $j \in \mathbb{N}_0$. Using reconstruction length $W \in \mathbb{N}$, we define a sliding sample $\vec{\xi}_j = (y_j, \ldots, y_{j+W-1}) \in \mathbb{R}^W$ for $j \in \mathbb{N}_0$, first. Whitney's embedding theorem can be rewritten from continuous to discrete time as follows: When $W \geq 2N + 1$, then the reconstructed series $\{\vec{\xi}_j\}_{j=0}^{\infty}$ has the same properties as $\{\vec{x}_j\}_{j=0}^{\infty}$. Therefore, any Renyi dimension D_α of the reconstructed attractor is the same as in the case of the state space.

Table 4 shows the comparison of the dimension estimation using the box-counting (denoted as box-count) method and the new method of the modified Renyi entropy (denoted as m. Renyi).

Table 4
Discrete dynamical system analysis.

system	α	D_α	data	method	$\widehat{D_\alpha}$	sd	p-value
Henon map	1	1.2583	OD	m. Renyi	1.2608	0.0156	**0.4363**
				box-count	1.2428	0.0113	**0.0851**
			RD	m. Renyi	1.2590	0.0056	**0.4503**
				box-count	1.2489	0.0031	0.0012
Henon map	2	1.2201	OD	m. Renyi	1.2243	0.0174	**0.4046**
				box-count	1.2161	0.0109	**0.3568**
			RD	m. Renyi	1.2230	0.0026	**0.1323**
				box-count	1.2172	0.0014	0.0192
Lozi map	1	1.4042	OD	m. Renyi	1.4131	0.0197	**0.3257**
				box-count	1.3915	0.0174	**0.2327**
			RD	m. Renyi	1.4098	0.0044	**0.1016**
				box-count	1.3945	0.0032	0.0012
Lozi map	2	1.3845	OD	m. Renyi	1.3937	0.0144	**0.2614**
				box-count	1.3786	0.0161	**0.3570**
			RD	m. Renyi	1.3885	0.0031	**0.0985**
				box-count	1.3749	0.0041	0.0096

The comparison has been performed for original state data (OD) and reconstructed data (RD). For the experiment, the Henon map [26,27] with the parameters $a = 0.4$, $b = 0.3$ and the starting points $x_{0,1} = 0, x_{0,2} = 0.9$ and the Lozi map [28,29] with the parameters $a = 1.7$, $b = 0.5$ and the starting points $x_{0,1} = -0.1, x_{0,2} = 0.1$ were used for the simulation for $\alpha \in \{1, 2\}$, reconstruction length $W = 5$, range for modified Renyi method as $f_{\min} = -2.0$ and $f_{\max} = -2.0$ and for all experiments, the set Φ contained $M = 10^6$ elements. The experiment was also conducted for bigger lengths of the reconstruction window, but it didn't have a significant impact on the results and their precision.

When the systems are investigated in the state space (OD) of low dimension ($n = 2$), the box-counting offered more accurate estimates with smaller standard deviation than the novel method. However, the p-values indicate unbiasedness in both cases. Another behaviour of estimation methods has been observed in the case of state reconstruction (RD) when the space dimension is large ($n = 5$). Therefore, the box-counting estimates of event probabilities are biased due to data sparsity. As seen in Table 4, all the box-counting estimates after reconstruction are biased. The sparsity effect is not present in the case of new method, where the p-values are higher with similar standard deviation. Therefore, the modified Renyi dimension is more suitable for reconstructed systems in higher-dimensional space, where the unbiasedness is present and the estimation accuracy is higher.

8. Conclusion

The paper presents new term modified entropy that has been defined using the Parzen formula with a ball kernel. The new entropy measure can be calculated for all finite samples Φ using a degeneracy function. The Monte Carlo approach enables the estimation of the proposed modified entropy which is later useful for the dimension estimation. The relationship between D_0, D_2 and their estimates from the modified entropy have been both theoretically and numerically proven for an arbitrary measurable set $\mathcal{F}$. Moreover, numerical simulations on selected fractal sets verified the unbiasedness of the D^*_α estimates.

Acknowledgement

The paper has been created with the support of CTU in Prague, Grant SGS14/208/OHK4/3T/14.

References

[1] Falconer K. Fractal geometry: mathematical foundations and applications. Wiley; 2014. ISBN 111994239X.
[2] Kirkby MJ. The fractal geometry of nature. Earth Surf Processes Landforms 1983;8(4) 406–406. doi:10.1002/esp.3290080415.
[3] Edgar G. Measure, topology, and fractal geometry (undergraduate texts in mathematics). Springer; 2007. ISBN 0387747486.
[4] Renyi A. Probability theory (north-Holland series in applied mathematics AND mechanics). Elsevier Science Publishing Co Inc.,U.S.; 1970. ISBN 0720423600.
[5] Sarkar N, Chaudhuri BB. An efficient differential box-counting approach to compute fractal dimension of image. IEEE Trans Syst Man Cybern 1994;24(1):115–20. doi:10.1109/21.259692.
[6] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D 1983;9(1–2):189–208. doi:10.1016/0167-2789(83)90298-1.
[7] Grassberger P, Procaccia I. Characterization of strange attractors. Phys Rev Lett 1983;50(5):346–9. doi:10.1103/physrevlett.50.346.
[8] Dlask M, Kukal J. Application of rotational spectrum for correlation dimension estimation. Chaos Solitons Fractals 2017;99:256–62. doi:10.1016/j.chaos.2017.04.026.
[9] Grassberger P. Finite sample corrections to entropy and dimension estimates. Phys Lett A 1988;128(6–7):369–73. doi:10.1016/0375-9601(88)90193-4.
[10] Smith LA. Intrinsic limits on dimension calculations. Phys Lett A 1988;133(6):283–8. doi:10.1016/0375-9601(88)90445-8.
[11] Yan J, Sun YY, Cai SS, Hu XP. An improved box-counting method to estimate fractal dimension of images. J Appl Anal Comput 2016;6(4):1114–25. doi:10.11948/2016073.
[12] Liu Y, Chen L, Wang H, Jiang L, Zhang Y, Zhao J, Wang D, Zhao Y, Song Y. An improved differential box-counting method to estimate fractal dimensions of gray-level images. J Vis Commun Image Represent 2014;25(5):1102–11. doi:10.1016/j.jvcir.2014.03.008.
[13] Sadowski W. A remark on the box-counting dimension of the singular set for the Navier-Stokes equations. Commun Math Sci 2013;11(2):597–602.
[14] Wang Y, Wu G. On the box-counting dimension of the potential singular set for suitable weak solutions to the 3d Navier–Stokes equations. Nonlinearity 2017;30(5):1762–72. doi:10.1088/1361-6544/aa6444.
[15] Shayegh F, Sadri S, Amirfattahi R, Ansari-Asl K. A model-based method for computation of correlation dimension, lyapunov exponents and synchronization from depth-EEG signals. Comput Methods Programs Biomed 2014;113(1):323–37. doi:10.1016/j.cmpb.2013.08.014.
[16] Rawal K, Saini BS, Saini I. Adaptive correlation dimension method for analysing heart rate variability during the menstrual cycle. Australas. Physical Eng Sci Med 2015;38(3):509–23. doi:10.1007/s13246-015-0369-y.
[17] Kalauzi A, Vuckovic A, Bojić T. Topographic distribution of EEG alpha attractor correlation dimension values in wake and drowsy states in humans. Int J Psychophysiol 2015;95(3):278–91. doi:10.1016/j.ijpsycho.2014.11.008.
[18] Nie C-X. Correlation dimension of financial market. Physica A 2017;473:632–9. doi:10.1016/j.physa.2017.01.045.
[19] Yue-chun Z, Guang-shui L. The research on determinants of capital structure based on fractal dimensionality reduction. In: Hua L, editor. 2012 International conference on management science and engineering. International Conference on Management Science and Engineering-Annual Conference Proceedings; IEEE; Univ Texas, Sch Management; Harbin Inst Technol, Sch Management; IEEE Technol Management Council; Natl Nat Sci Fdn China (NSFC); Harbin Inst Technol; 2012. p. 1553–9. ISBN 978-1-4673-3015-2; 978-1-4673-3014-5; 19th Annual International Conference on Management Science and Engineering (ICMSE), Dallas, TX, SEP 20–22, 2012.
[20] Parzen E. On estimation of a probability density function and mode. Annals Math Stat 1962;33(3):1065–76. doi:10.1214/aoms/1177704472.
[21] OLSON EJ, ROBINSON JC, SHARPLES N. Generalised cantor sets and the dimension of products. Math Proc Cambridge Philos Soc 2015;160(01):51–75. doi:10.1017/s0305004115000584.
[22] Bandt C, Hung NV, Rao H. On the open set condition for self-similar fractals. Proc Am Math Soc 2005;134(05):1369–74. doi:10.1090/s0002-9939-05-08300-0.
[23] Takayasu H. Fractals in the physical sciences (nonlinear science). Manchester Univ Pr; 1990. ISBN 0719024854.
[24] Rajković N, Krstonošić B, Milošević N. Box-counting method of 2d neuronal image: method modification and quantitative analysis demonstrated on images from the monkey and human brain. Comput Math Methods Med 2017;2017:1–9. doi:10.1155/2017/8967902.
[25] Whitney H. Differentiable manifolds. Annals Math 1936;37(3):645. doi:10.2307/1968482.
[26] Cao Y-L. Strange attractor of henon map and its basin. Science in China. Series A 1995;38.
[27] Hénon M. A two-dimensional mapping with a strange attractor. Comm Math Phys 1976;50(1):69–77.
[28] Misiurewicz M, Štimac S. Symbolic dynamics for lozi maps. Nonlinearity 2016;29(10):3031–46. doi:10.1088/0951-7715/29/10/3031.
[29] Lozi R. Un attracteur etrange du type attracteur de henon. Le J de Physique Colloques 1978;39(C5) C5–9–C5–10. doi:10.1051/jphyscol:1978505.

Physica A 585 (2022) 126424

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Hurst exponent estimation of fractional surfaces for mammogram images analysis

Martin Dlask *, Jaromir Kukal

Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Trojanova 13, Praha, Czech Republic

ARTICLE INFO

Article history:
Received 2 January 2021
Received in revised form 7 July 2021
Available online 15 September 2021

Keywords:
Fractional Brownian motion
Hurst exponent
Wishart distribution
Mammogram images

ABSTRACT

The work presents a methodology to precise simulation and parameter estimation of multidimensional fractional Brownian motion (fBm). The simulation approach uses circulant embedding algorithm and solution of Poisson equation, while generalizing it to multiple dimensions. For estimation, a method using Wishart distribution and maximum likelihood is presented and verified on simulated data. Unlike approximate methods for generating multidimensional fBm and its Hurst exponent estimation, this approach shows unbiased results for all processes with short memory and majority of cases with long memory. The methodology is applied to mammography screening images to find significant differences between benign and cancerous breast lumps.

1. Introduction

Fractional Brownian motion (fBm) introduced first in [1] is a stochastic process often used for modeling processes with long or short memory [2]. Having fractional character depending on parameter $H \in (0, 1)$ that is often called Hurst exponent [3], the process is continuous and has stationary increments. The trajectory of fBm signal in one dimension has fractal character and its fractal dimension [4] equals $2 - H$. This process is advantageously used for modeling of stochastic events, especially in biomedicine and economics.

Many real world phenomena, however, do not occur only in the shape of random processes, but in the shape of random surfaces or random mass. There is a variety of methods that can analyze sets in multiple dimension, such as variations of box-counting algorithm [5] and algorithms for point sets, such as rotational spectrum [6] or entropy estimates [7]. By terms of random processes in higher dimensional spaces and without the loss of generality, the multidimensional fBm [8] can be defined via extending its definition from one dimension. This paper describes the basic properties of multivariate fBm and presents the way of estimating its Hurst exponent.

The samples of multidimensional fBm can be simulated using variety of methods. The first category of methods, which is often used for terrain modeling, is using technique of random point displacement [9]. This kind of methods is very simple to implement, however, the samples do not conform with the prescribed autocorrelation function. An exact approach involves the generation using Cholesky decomposition [10] and can provide very accurate samples of fractional Brownian surfaces and masses. Its major disadvantage is its computational demand, which makes it unusable for simulations of larger sets. The methods that compromise between the accuracy and speed and in theory converge to the autocorrelation structure of multidimensional fBm leverage the principle of generalizing fractional Gaussian processes [11] or circulant embedding [12].

* Corresponding author.
E-mail address: martindlask@centrum.cz (M. Dlask).

https://doi.org/10.1016/j.physa.2021.126424
0378-4371/

The idea of analyzing multidimensional fBm using the separation of two-dimensional trajectories into one-dimensional ones have been proven to be useful for Bayesian analysis of single-particle tracking data [13] as well as in diffusion models [14].

Accurate estimation of Hurst exponent can be beneficial for other research areas that investigate samples with chaotic character. As an example, the human brain [15,16] can be the subject of fractal investigation in the neuroscience area with aim to discover early stages of Alzheimer disease. Another application area is astronomy, where the structure of star clusters can be efficiently analyzed using multivariate fBm [17]. The mammography images [18] can be the subject of analysis with respect to the detection of benign and cancerous lumps.

2. Multidimensional fractional Brownian motion

The one-dimensional fractional Brownian motion (fBm) with Hurst exponent $H \in (0, 1)$ and variance $\sigma_0^2 > 0$ can be generalized as follows [8,19]. Let $\vec{x} \in \mathbb{R}^n$ be independent variable of dimension $n \in \mathbb{N}$. The n-dimensional fBm [20] denoted as $B_H(\vec{x})$ satisfies

$$\mathrm{E}\left(B_H(\vec{x})B_H(\vec{y})\right) = \frac{\sigma_0^2}{2} \cdot \left(\|\vec{x}\|_2^{2H} + \|\vec{y}\|_2^{2H} - \|\vec{x} - \vec{y}\|_2^{2H}\right) \tag{1}$$

and $\mathrm{E}\, B_H(\vec{x}) = 0$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ where $\|..\|_2$ denotes Euclidean distance. Using rectangular equidistant sampling with step $h > 0$ and denoting $B(\vec{K}) = B_H(h\vec{K})$ for $\vec{K} \in \mathbb{Z}^n$, we obtain discrete fBm formula satisfying $\mathrm{E}\, B(\vec{K}) = 0$ and

$$\mathrm{E}\left(B(\vec{K})B(\vec{L})\right) = \frac{\sigma^2}{2}\left(\|\vec{K}\|_2^{2H} + \|\vec{L}\|_2^{2H} - \|\vec{K} - \vec{L}\|_2^{2H}\right) \tag{2}$$

for $\vec{K}, \vec{L} \in \mathbb{Z}^n$, where $\sigma = \sigma_0 \cdot h^H$.

The discretized n-dimensional signal can be sampled by symmetric compact mask

$$\mathcal{M}^* = \left\{\vec{s} \in \mathbb{Z}^n : \|\vec{s}\|_p \le \rho\right\} \tag{3}$$

using Minkowski norm $\|..\|_p$ for $p, \rho \ge 1$. The parameter ρ is later called mask radius. The mask $\mathcal{M}^*$ consists of $N + 1$ points $\vec{s}_0 = \vec{0}, \vec{s}_1, \ldots, \vec{s}_N$ and can be shifted from the origin to a point $\vec{K} \in \mathbb{Z}^n$ to form

$$\mathcal{M}(\vec{K}) = \left\{\vec{K} + \vec{s} : \vec{s} \in \mathcal{M}^*\right\}. \tag{4}$$

Therefore, the sample at point $\vec{K}$ consists of values $B(\vec{K}), B(\vec{K} + \vec{s}_1), \ldots, B(\vec{K} + \vec{s}_N)$. The neighborhood of $B(\vec{K})$ and its generation for rectangular mask with nine elements is visualized in Fig. 1. We investigate the reduced sample of size N and values $\xi_i = B(\vec{K} + \vec{\xi}_i) - B(\vec{K})$ for $i = 1, \ldots, N$. Adequate covariance matrix $\mathbb{V} = \mathrm{var}\, \vec{x}_i$ has elements

$$v_{i,j} = \mathrm{E}(\xi_i \xi_j) = \mathrm{E}\left(B(\vec{K} + \vec{s}_i) - B(\vec{K})\right)\left(B(\vec{K} + \vec{s}_j) - B(\vec{K})\right) \tag{5}$$

$$= \mathrm{E}\left(B(\vec{s}_i) - B(\vec{0})\right)\left(B(\vec{s}_j) - B(\vec{0})\right) \tag{6}$$

$$= \frac{\sigma^2}{2} \cdot \left(\|\vec{s}_i\|_2^{2H} + \|\vec{s}_j\|_2^{2H} - \|\vec{s}_i - \vec{s}_j\|_2^{2H}\right) \tag{7}$$

which is independent on $\vec{K}$. The theoretical covariance matrix of the sample has the same dependences as the covariance matrix of the original process.

3. Method of estimation

Given n-dimensional fBm and mask $\mathcal{M}^*$, we consider m sets $\mathcal{M}(\vec{r}_q)$ where $\vec{r}_q \in \mathbb{Z}^n$ for $q = 1, \ldots, m$. The parameter m will be further called number of samples. For each set, we define sample vector $\vec{g}_q \in \mathbb{R}^N$ as

$$\vec{g}_q^{\mathrm{T}} = \left(B(\vec{r}_q + \vec{s}_1) - B(\vec{r}_q), B(\vec{r}_q + \vec{s}_2) - B(\vec{r}_q), \ldots, B(\vec{r}_q + \vec{s}_N) - B(\vec{r}_q)\right). \tag{8}$$

for $q = 1, \ldots, m$. Let $\mathbb{G} = \left(\vec{g}_1, \vec{g}_2, \ldots, \vec{g}_m\right) \in \mathbb{R}^{N \times m}$ be the sample matrix, having $\vec{g}_q \sim \mathrm{N}(\vec{0}, \mathbb{V})$. The scatter matrix $\mathbb{S} = \mathbb{G}\mathbb{G}^{\mathrm{T}}$ has Wishart distribution [21] with probability density function

$$f(\mathbb{S}, \mathbb{V}, N, m) = \frac{|\mathbb{S}|^{\frac{m-N-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbb{V}^{-1}\mathbb{S})\right)}{2^{mN/2}\, |\mathbb{V}|^{m/2}\, \Gamma_N(\frac{m}{2})}, \tag{9}$$

where

$$\Gamma_N(\xi) = \pi^{N(N-1)/4} \cdot \prod_{j=1}^{N} \Gamma\left(\xi - \frac{j-1}{2}\right) \tag{10}$$

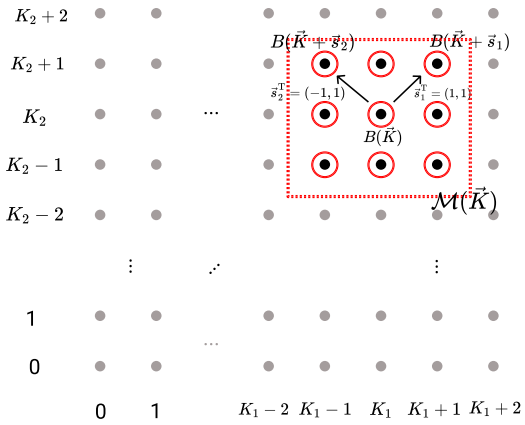


Fig. 1. Example of rectangular mask applied to two-dimensional fBm.

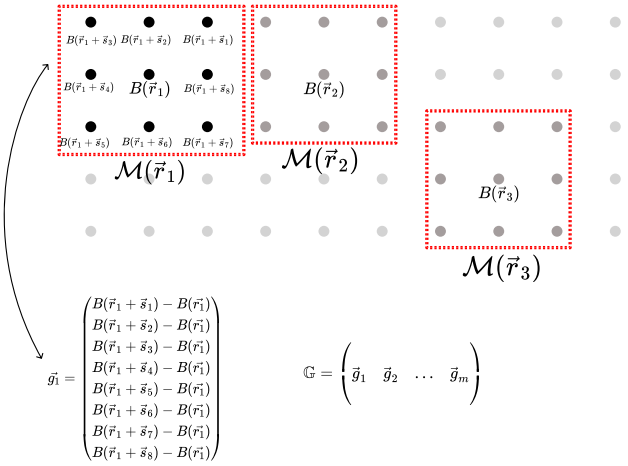


Fig. 2. Scatter matrix generation from two-dimensional fBm sample.

and tr(...) means trace of a matrix.

Given $m > N$ samples of n-dimensional fBm, one can calculate the scatter matrix as

$$\mathbb{S} = \sum_{l=1}^{m} \vec{g}_j \vec{g}_j^{\mathrm{T}}. \tag{11}$$

The scatter matrix generation from two-dimensional fBm and rectangular mask is visualized in Fig. 2.

By maximizing the value of corresponding likelihood

$$L(H, \sigma) = f(\mathbb{S}, \mathbb{V}(H, \sigma), N, m) \to \max, \tag{12}$$

one gets the optimum values (H^*, σ^*) along with the standard deviations of the estimates.

4. Generalization of circulant embedding method

The idea of novel multidimensional fBm generation is based on the generalization of traditional circulant embedding [22,23] (TCE) method. Now we introduce general circulant embedding (GCE) method as follows. The method utilizes the simulation of incremental fractional Gaussian noise (ifGn) signal in its core. Suppose that M is even size of n-dimensional integer grid $\mathcal{D} = \{-M/2, \ldots, M/2-1\}^n$ which is a support of $B_H(\vec{k})$ as fBm and $Y_H(\vec{k})$ as ifGn for $\vec{k} \in \mathcal{D}$.

Using index vector $\vec{k} \in \mathcal{D}$, we will study the relationship between $B_H(\vec{k})$ and $Y_H(\vec{k})$. Using discrete Laplacian Δ_D in cubic domain, we can directly calculate

$$Y_H(\vec{k}) = \Delta_D B_H(\vec{k}) = \sum_{\substack{\vec{u}\in\mathbb{Z}^n \\ \|\vec{u}\|=1}} \left(B_H(\vec{k}+\vec{u}) - B_H(\vec{k}) \right). \tag{13}$$

Having $B_H(\vec{0}) = 0$, we can continue with

$$C(\vec{k}) = \mathrm{E}\left(Y_H((\vec{k}) Y_H(\vec{0})) \right) \tag{14}$$

$$= \mathrm{E}\left(\sum_{\substack{\vec{u}\in\mathbb{Z}^n \\ \|\vec{u}\|=1}} \left(B_H(\vec{k}+\vec{u}) - B_H(\vec{k}) \right) \cdot \sum_{\substack{\vec{v}\in\mathbb{Z}^n \\ \|\vec{v}\|=1}} \left(B_H(\vec{v}) - B_H(\vec{0}) \right) \right) \tag{15}$$

$$= \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \mathrm{E}\left(B_H(\vec{k}+\vec{u}) B_H(\vec{v}) - B_H(\vec{k}) B_H(\vec{v}) \right) \tag{16}$$

$$= \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \mathrm{E}\left(B_H(\vec{k}+\vec{u}) B_H(-\vec{v}) - B_H(\vec{k}) B_H(-\vec{v}) \right) \tag{17}$$

$$= \frac{\sigma^2}{2} \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \Big(\|\vec{k}+\vec{u}\|^{2H} + \| - \vec{v}\|^{2H} - \|\vec{k}+\vec{u}+\vec{v}\|^{2H} - \|\vec{k}\|^{2H} - \| - \vec{v}\|^{2H} + \|\vec{k}+\vec{v}\|^{2H} \Big) = \tag{18}$$

$$= \frac{\sigma^2}{2} \sum_{\substack{\vec{u},\vec{v}\in\mathbb{Z}^n \\ \|\vec{u}\|=\|\vec{v}\|=1}} \left(\|\vec{k}\|^{2H} - \|\vec{k}+\vec{u}\|^{2H} - \|\vec{k}+\vec{v}\|^{2H} + \|\vec{k}+\vec{u}+\vec{v}\|^{2H} \right). \tag{19}$$

Using matrix $\mathbb{W}_H \in \mathbb{R}^{M^n}$ of elements $W_H(\vec{k}) = \|\vec{k}\|^{2H}$ for $\vec{k} \in \mathcal{D}$ we can also apply the discrete Laplacian Δ_D to produce

$$\Delta_D W_H(\vec{k}) = \sum_{\substack{\vec{u}\in\mathbb{Z}^n \\ \|\vec{u}\|=1}} \left(W_H(\vec{k}+\vec{u}) - W_H(\vec{k}) \right) \tag{20}$$

and also the convolution

$$\left(\Delta_D * \Delta_D \right) W_H(\vec{k}) = \tag{21}$$

$$= \sum_{\substack{\vec{v}\in\mathbb{Z}^n \\ \|\vec{v}\|=1}} \left(\Delta_D W_H(\vec{k}+\vec{v}) - \Delta_D W_H(\vec{k}) \right) = \tag{22}$$

$$= \sum_{\substack{\vec{v},\vec{u}\in\mathbb{Z}^n \\ \|\vec{v}\|=\|\vec{u}\|=1}} \left(W_H(\vec{k}+\vec{u}+\vec{v}) - W_H(\vec{k}+\vec{u}) - W_H(\vec{k}+\vec{v}) + W_H(\vec{k}) \right) = \tag{23}$$

$$= \sum_{\substack{\vec{v},\vec{u}\in\mathbb{Z}^n \\ \|\vec{v}\|=\|\vec{u}\|=1}} \left(\|\vec{k}+\vec{u}+\vec{v}\|^{2H} - \|\vec{k}+\vec{u}\|^{2H} - \|\vec{k}+\vec{v}\|^{2H} + \|\vec{k}\|^{2H} \right). \tag{24}$$

Therefore,

$$C(\vec{k}) = -\frac{\sigma^2}{2} \left(\Delta_D * \Delta_D \right) W_H(\vec{k}) \tag{25}$$

or in matrix form as

$$\mathbb{C} = -\frac{\sigma^2}{2} \left(\Delta_D * \Delta_D \right) \mathbb{W}_H. \tag{26}$$

Process Y_H is still Gaussian with known autocovariance structure. Moreover, its autocorrelation is not dependent on the time lag, therefore it can be generated using the circulant embedding method.

For instance in case $n = 1$, the process Y_H equals

$$Y_H(j) = B_H(j+1) - 2B_H(j) + B_H(j-1), \tag{27}$$

in case of $n = 2$ the process formula can be expressed as

$$Y_H(j_1, j_2) = B_H(j_1 + 1, j_2) + B_H(j_1, j_2 + 1) + \\ + B_H(j_1 - 1, j_2) + B_H(j_1, j_2 - 1) - 4B_H(j_1, j_2). \tag{28}$$

Having randomly generated sample of ifGn as $Y_H(\vec{k})$ with zero mean value, we recognize (13) as Discrete Poisson Equation [24] for an instance of fBm $B_H(\vec{k})$ on the integer domain $\mathcal{D}$. Denoting DFT, DFT^{-1} as Discrete Fourier transform and its inversion on $\mathcal{D}$, we can directly solve (13) with periodic boundary conditions. In this case, (13) has infinite many solutions which differ by constants. Without loosing generality, we will find the solution with zero mean value as follows. First, we will apply DFT to (13) and obtain

$$\mathrm{DFT}\{Y_H(\vec{k})\} = F(\vec{k}) \otimes \mathrm{DFT}\{B_H(\vec{k})\},$$

where

$$F(\vec{k}) = -4 \sum_{j=1}^{n} \sin^2(k_j/2) \in \mathbb{R}$$

for $\vec{k} \in \mathcal{D}$ and $\otimes$ is component-wise multiplication. The deconvolution can be obtained using Wiener filter [25,26] with parameter $\lambda > 0$ using formula

$$B_H(\vec{k}) = \mathrm{DFT}^{-1}\{W(\vec{k}) \otimes \mathrm{DFT}\{Y_H(\vec{k})\}\}, \tag{29}$$

where

$$W(\vec{k}) = \frac{F^*(\vec{k})}{\lambda^2 + |F(\vec{k})|^2} = \frac{F(\vec{k})}{\lambda^2 + F^2(\vec{k})}. \tag{30}$$

5. Application to simulated data

The method of Hurst exponent estimation and its standard deviation is first validated on simulated data. At first, we investigated the optimal value of regularization parameter $\lambda > 0$ in one and two dimensions. It has been observed the Wiener method of regularization numerically fails for $\lambda < 10^{-13}$ in many cases. Therefore, we suggest $\lambda = 10^{-12}$ as the minimum but safe value.

Using this value we can generate the n-dimensional fBm and estimate H using the maximum likelihood [27] method and Wishart approach. The experiment has been performed in one and two dimensions. The theoretical value of H is uniformly generated from 0.05 to 0.95.

The results for one dimensional case (traditional fractional Brownian motion) for TCE and GCE methods are shown in Table 1. While H_{theory} denotes the theoretical value of Hurst exponent, $\widehat{H}$ is its estimate with corresponding standard deviation s. The column z denotes the z-score as a measure of the unbiasedness of the estimate. Mean square error (MSE) is defined as

$$\mathrm{MSE} = (H_{\mathrm{theory}} - \widehat{H})^2 + s^2. \tag{31}$$

Absolute value of z less than 1.96 cannot reject the null hypothesis on critical level 0.05, that the estimated Hurst exponent from the sample would significantly differ from its theoretical value. In our case this is fulfilled for all cases except $H = 0.95$, when the fractional Brownian motion has a very long memory and both methods overestimate its theoretical value. Additionally, in all cases when the estimates are unbiased, the novel GCE method is having smaller MSE. The performance of TCE method is visualized in Fig. 3.

In two dimensional space, where the equivalent of TCE does not exist, the experiment is performed for GCE and symmetric compact mask with $p = 1$ and $\rho = 4$, yielding $N = 80$. The estimation has been performed from $m = 10^4$ independent samples and the results are presented in Table 2. The method provided unbiased estimation of Hurst exponent for all $H < 0.90$, except the extreme case of long memory fractional Brownian surface. The performance of the GCE method is visualized in Fig. 4.

6. Application to mammography

The theory of Wishart distribution and Hurst exponent estimation from sample is applied to the data from mammography screening, assuming the scan of woman's breast is a sample of fractional Brownian surface. The data are taken from the Digital Database for Screening Mammography [28,29] containing public and open-access data for research in image processing. For the study, we have selected 140 patients who have benign lump (BL) and 140 patients with cancerous lump (CL). The cranio caudal (CC) images are used for the purposes of this study. (See Figs. 5 and 6)

The images were captured by DBA M2100 ImageClear digitizer with sampling rate 42 μm and 16 bits gray level. The pictures were processed in Matlab as grayscale pictures and cropped, so the picture only contains human tissue without background noise.

Table 1
Hurst exponent estimation of fBm in one dimension for $N = 10$ and $m = 10^4$.

H_{theory}	TCE				GCE			
	$\widehat{H}$	s	z	MSE	$\widehat{H}$	s	z	MSE
0.0500	0.0459	0.00276	1.49	$2.46 \cdot 10^{-5}$	0.0498	0.00141	0.12	$2.03 \cdot 10^{-6}$
0.1000	0.1010	0.00354	−0.29	$1.36 \cdot 10^{-5}$	0.0997	0.00180	0.19	$3.33 \cdot 10^{-6}$
0.1500	0.1519	0.00405	−0.47	$2.00 \cdot 10^{-5}$	0.1538	0.00203	−1.85	$1.86 \cdot 10^{-5}$
0.2000	0.1923	0.00440	1.75	$7.84 \cdot 10^{-5}$	0.2020	0.00222	−0.88	$8.93 \cdot 10^{-6}$
0.2500	0.2578	0.00472	−1.65	$8.24 \cdot 10^{-5}$	0.2479	0.00236	0.90	$9.98 \cdot 10^{-6}$
0.3000	0.3013	0.00480	−0.26	$2.46 \cdot 10^{-5}$	0.3033	0.00246	−1.35	$1.69 \cdot 10^{-5}$
0.3500	0.3509	0.00502	−0.18	$2.60 \cdot 10^{-5}$	0.3487	0.00248	0.52	$7.84 \cdot 10^{-6}$
0.4000	0.4068	0.00504	−1.34	$7.10 \cdot 10^{-5}$	0.4006	0.00252	−0.23	$6.71 \cdot 10^{-6}$
0.4500	0.4482	0.00493	0.36	$2.74 \cdot 10^{-5}$	0.4494	0.00256	0.22	$6.91 \cdot 10^{-6}$
0.5000	0.4956	0.00509	0.87	$4.54 \cdot 10^{-5}$	0.5002	0.00250	−0.09	$6.29 \cdot 10^{-6}$
0.5500	0.5495	0.00497	0.10	$2.50 \cdot 10^{-5}$	0.5512	0.00251	−0.46	$7.74 \cdot 10^{-6}$
0.6000	0.6001	0.00493	−0.01	$2.43 \cdot 10^{-5}$	0.6024	0.00249	−0.96	$1.20 \cdot 10^{-5}$
0.6500	0.6515	0.00477	−0.32	$2.51 \cdot 10^{-5}$	0.6501	0.00237	−0.04	$5.63 \cdot 10^{-6}$
0.7000	0.6942	0.00457	1.27	$5.48 \cdot 10^{-5}$	0.7002	0.00222	−0.07	$4.97 \cdot 10^{-6}$
0.7500	0.7459	0.00440	0.93	$3.63 \cdot 10^{-5}$	0.7490	0.00218	0.44	$5.75 \cdot 10^{-6}$
0.8000	0.8020	0.00397	−0.50	$1.97 \cdot 10^{-5}$	0.7989	0.00197	0.57	$5.09 \cdot 10^{-6}$
0.8500	0.8531	0.00346	−0.90	$2.16 \cdot 10^{-5}$	0.8490	0.00175	0.57	$4.06 \cdot 10^{-6}$
0.9000	0.9052	0.00267	−1.95	$3.44 \cdot 10^{-5}$	0.8979	0.00147	1.45	$6.57 \cdot 10^{-6}$
0.9500	0.9444	0.00175	3.22	$3.49 \cdot 10^{-5}$	0.9594	0.00088	−10.67	$8.91 \cdot 10^{-5}$

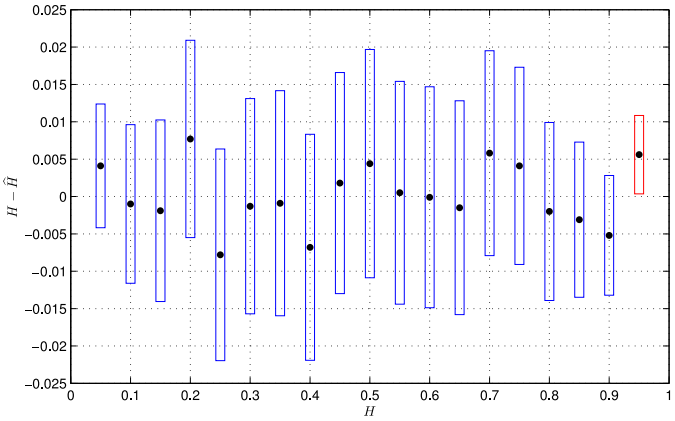


Fig. 3. TCE method performance in one-dimensional space.

Afterwards, the threshold technique has been applied to the images. For image with dimensions $P, Q \in \mathbb{N}$, we define dimensionless parameter $\theta \in (0; 1)$ and define our region of interest as

$$ROI = \{(i, j) : x_{i,j} > \theta \cdot \max_{\substack{1 \le i \le Q \\ 1 \le j \le P}} x_{i,j}\} \tag{32}$$

to localize the benign and cancerous lumps on the breast image. For our experiment, we have chosen three thresholds selecting points with high intensity from the original x-ray scans. The threshold $\theta = 0.80$ corresponds to roughly 5%, threshold $\theta = 0.85$ assures roughly 3% and threshold $\theta = 0.9$ corresponds to roughly 1% of pixels of the original image. For generating scatter matrix, we will use symmetric masks with radius from three to seven pixels, i.e. $\rho \in \{3, 4, 5, 6, 7\}$.

We will use two approaches to compare the results between BL and CL patients. First approach estimates H for each patient in the respective groups. Using Wilcoxon ranksum test, we will test hypothesis

$$H_0 : H_{\text{B}} = H_{\text{C}}, \tag{33}$$

where H_{B} and H_{C} are the Hurst exponent estimates of BL and CL groups, respectively. The results of the estimation are displayed in Table 3. Using false discovery rate, we diminish the critical level of significance to $\alpha_{\text{FDR}} = 0.0217$. The cases, where the null hypothesis is rejected are highlighted in bold.

The case with highest significance (lowest p-value) in the traditional approach is $(\theta, \rho) = (0.85, 5)$. This combination will be investigated further using another approach that employs the estimation of both parameters (H, σ) and the

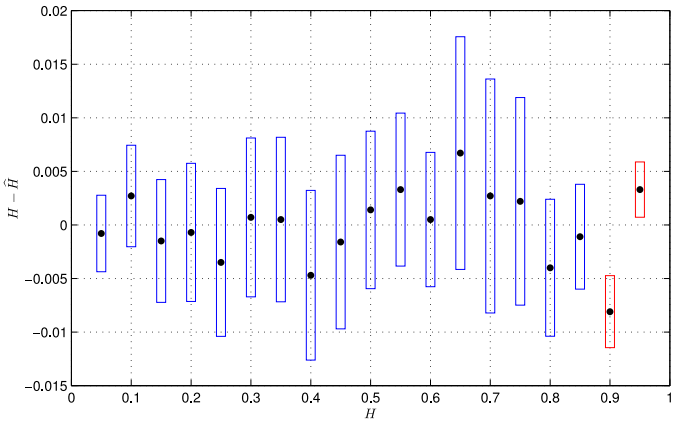


Fig. 4. GCE method performance in two-dimensional space.

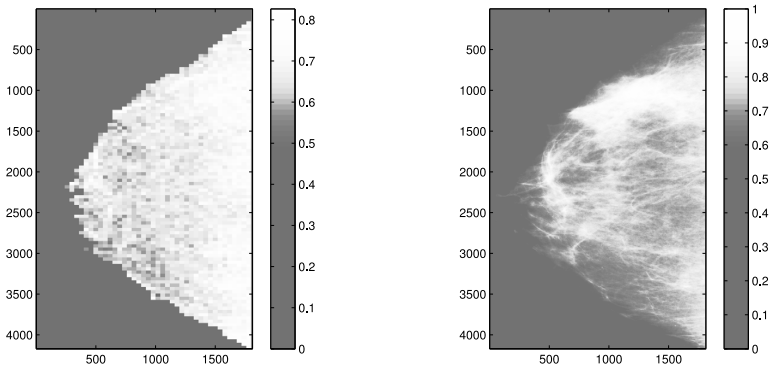


Fig. 5. Hurst exponent distribution map (left) and the original intensity image (right) of a CC image of woman's right breast.

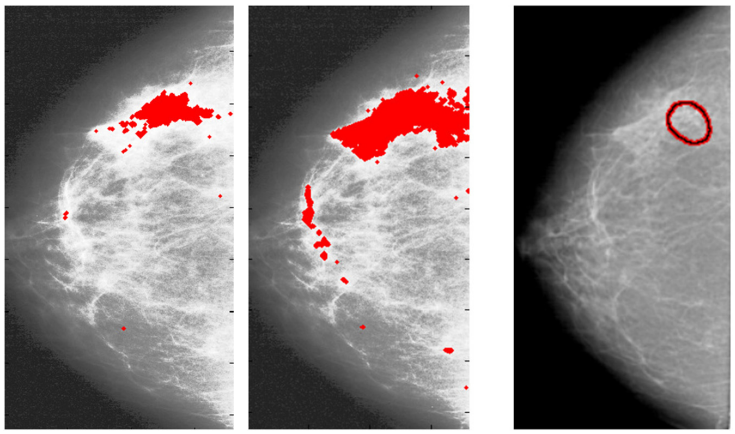


Fig. 6. Regions selected for analysis for benign finding for $\theta = 0.9$ (left), $\theta = 0.85$ (middle) and affected area highlighted by a medical doctor (right).

Table 2
Hurst exponent estimation of fBm in two dimensions for $N = 80$ and $m = 10^4$.

H_{theory}	$\widehat{H}$	s	z	MSE
0.0500	0.0508	0.00119	−0.66	$2.06 \cdot 10^{-6}$
0.1000	0.0973	0.00158	1.73	$9.79 \cdot 10^{-6}$
0.1500	0.1515	0.00191	−0.77	$5.90 \cdot 10^{-6}$
0.2000	0.2007	0.00215	−0.35	$5.11 \cdot 10^{-6}$
0.2500	0.2535	0.00230	−1.54	$1.75 \cdot 10^{-5}$
0.3000	0.2993	0.00247	0.29	$6.59 \cdot 10^{-6}$
0.3500	0.3495	0.00256	0.20	$6.80 \cdot 10^{-6}$
0.4000	0.4047	0.00264	−1.80	$2.91 \cdot 10^{-5}$
0.4500	0.4516	0.00270	−0.60	$9.85 \cdot 10^{-6}$
0.5000	0.4986	0.00845	0.17	$7.34 \cdot 10^{-5}$
0.5500	0.5467	0.00238	1.39	$1.66 \cdot 10^{-5}$
0.6000	0.5995	0.00209	0.24	$4.62 \cdot 10^{-6}$
0.6500	0.6433	0.00362	1.85	$5.80 \cdot 10^{-5}$
0.7000	0.6973	0.00364	0.75	$2.05 \cdot 10^{-5}$
0.7500	0.7478	0.00323	0.68	$1.53 \cdot 10^{-5}$
0.8000	0.8040	0.00213	−1.87	$2.05 \cdot 10^{-5}$
0.8500	0.8511	0.00163	−0.68	$3.87 \cdot 10^{-6}$
0.9000	0.9081	0.00112	−7.28	$6.69 \cdot 10^{-5}$
0.9500	0.9467	0.00086	3.82	$1.16 \cdot 10^{-5}$

Table 3
Wilcoxon–Mann–Whitney test for various thresholds and masks.

θ	ρ	H_{B}	std(H_{B})	H_{C}	std(H_{C})	p-val
0.8	3	0.5513	0.1912	0.5663	0.1891	0.1560
0.8	4	0.5943	0.1768	0.6281	0.1680	0.0732
0.8	5	0.5833	0.1766	0.6172	0.1733	0.0623
0.8	6	0.6384	0.1885	0.6664	0.1744	0.1386
0.8	7	0.6225	0.1866	0.6518	0.1737	0.1341
0.85	3	0.5313	0.1972	0.5782	0.2258	**0.0142**
0.85	4	0.6195	0.1752	0.6645	0.1927	**0.0107**
0.85	5	0.6158	0.1803	0.6601	0.1956	**0.0104**
0.85	6	0.6793	0.1917	0.7179	0.2029	**0.0192**
0.85	7	0.6616	0.1887	0.7018	0.2025	**0.0217**
0.9	3	0.4964	0.2454	0.5102	0.2781	0.0634
0.9	4	0.6551	0.1877	0.6908	0.2106	0.0245
0.9	5	0.6586	0.1895	0.6961	0.2199	**0.0212**
0.9	6	0.7300	0.2078	0.7549	0.2190	0.0445
0.9	7	0.7084	0.2053	0.7378	0.2178	0.0476

hypothesis claims, that the observed images of both cancerous and benign patterns are samples of fractional Brownian surface, but they can vary in parameters H and/or σ. We denote H_B as the Hurst exponent of the BL patients and H_C as the Hurst exponent of the CL patients and the corresponding σ as σ_C and σ_B respectively. Therefore, there are four sub-models in run:

1. model $H = H_B = H_C$ and $\sigma = \sigma_C = \sigma_B$
2. model $H = H_B = H_C$ and σ_C, σ_B
3. model H_B, H_C and $\sigma = \sigma_C = \sigma_B$
4. model H_B, H_C and σ_C, σ_B

Their hierarchy is depicted on Fig. 7. Both H and σ parameters can vary due to structural changes of the tissue. The estimated values of parameter σ also depend on X-ray intensity and sample depth, therefore their differences are of low practical importance. The estimated values of parameter H also depend on scanning apparatus resolution, which is constant for given data set and therefore more suitable for biomedical interpretation. The results of parameters for each of the models separately is presented in Table 4.

The models will be compared using likelihood ratio test. We can perform 4 comparisons as 2 to 1 (comparison D), 3 to 1 (comparison C), 4 to 2 (comparison B) and 4 to 3 (comparison A), while the first mentioned model in the pair is called full model and the latter one is its submodel. Having four comparisons in place allows us to measure, whether the difference between the BL and CL is in the Hurst exponent H or in the σ. For each of the models, we estimate its parameters by maximizing (12) and then calculate corresponding logarithm of likelihood as

$$\ln L = \ln f(\mathbb{S}, \mathbb{V}, p, m) \tag{34}$$

using (9). Comparing any model to its submodel, we calculate the log-likelihood of the full model and denote it $\ln L_{\text{FULL}}$, while the log-likelihood of the submodel equals $\ln L_{\text{SUB}}$. In our setting, any full model always differs in one free parameter

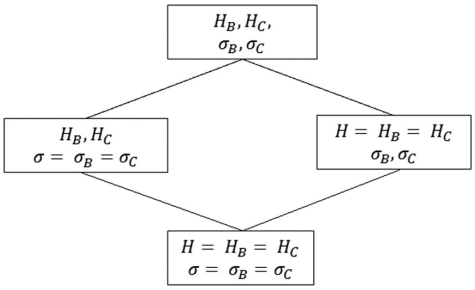


Fig. 7. Model hierarchy schema.

Table 4
Estimates of model parameters.

Model	H_B	H_C	σ_B	σ_C
1	0.5037	0.5037	0.0073	0.0073
2	0.5034	0.5043	0.0070	0.0075
3	0.5072	0.5085	0.0073	0.0073
4	0.4980	0.5085	0.0070	0.0075

Table 5
Model and sub-model comparison.

Full-model	Sub-model	Comparison	χ^2	$\ln p$-val
4	3	A	$1.13 \cdot 10^5$	$-5.63 \cdot 10^4$
4	2	B	$1.68 \cdot 10^3$	$-8.46 \cdot 10^2$
3	1	C	$1.53 \cdot 10^5$	$-7.63 \cdot 10^4$
2	1	D	$4.16 \cdot 10^4$	$-2.08 \cdot 10^4$

in comparison to its sub-model. It holds that

$$2\,(\ln L_{\text{FULL}} - \ln L_{\text{SUB}}) \sim \chi^2. \tag{35}$$

Using the likelihood ratio tests, the comparisons A,B,C,D were evaluated, the χ^2 value together with the logarithms of p-values were calculated. The comparisons are depicted in Table 5. All four sub models significantly differ in the sense of likelihood ratio test and therefore the fourth model i.e. the full model statistically dominates the others. It implies the rejection of hypothesis $H_B = H_C$.

Both Wilcoxon-approach combined with the false discovery rate as well as the likelihood ratio test have discovered significant differences between the CL and BL patients. The image of CL patients has statistically higher Hurst exponent when considering the tissue has properties of fractional Brownian surface. The distribution of Hurst exponent in case of CL and BL for $(\theta, \rho) = (0.85, 5)$ is visualized using boxplot in Fig. 8.

The CL patients have higher mean of Hurst exponent estimates, though their distribution significantly overlaps with BL patients as well. We have tested a naive-classifier that is using threshold

$$H_T = \frac{\mathrm{E}\,H_B + \mathrm{E}\,H_C}{2} \tag{36}$$

classifying any patient with $H > H_T$ as CL and others as BL. This naive classifier yields sensitivity 55% and accuracy 59% due to simple design of the classifier and due to the threshold technique does not always identify the affected tissues correctly. There have been other approaches that use combined information of Hurst exponent of the image, wavelet analysis and data distribution [30], yielding better performance in accuracy of 65% and sensitivity of 72%. Therefore, we recommend our algorithm for accurate H estimation that could serve as a powerful feature for image classification.

7. Conclusion

The paper presents findings from three areas — fBm generation, fBm estimation and fBm application to mammography.

The fBm simulation utilizes the generalization of the circulant embedding method. The idea of the signal generation is expanded to multiple dimensions and allows to generate the fractional Brownian surfaces. In contrast to TCE, the method uses the solution of discrete Poisson equation and Wiener filter with regularization parameter. This methodology was proven to be efficient in both one- and two-dimensional space for $\lambda = 10^{-12}$ and was shown to be at least the same efficient as TCE.

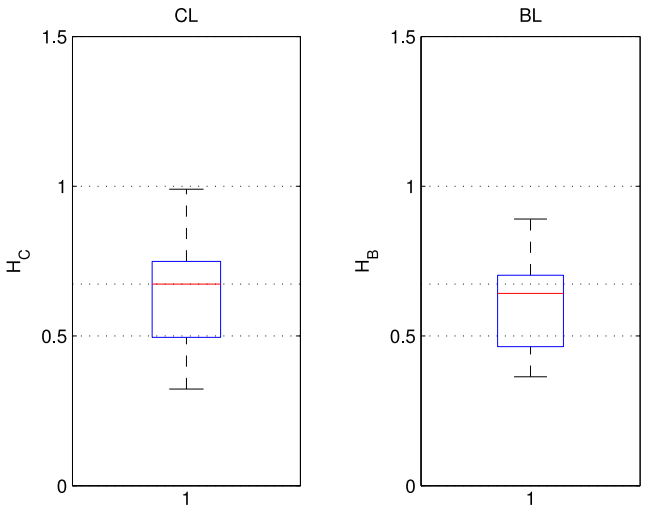


Fig. 8. Distribution of Hurst exponent in case of BL (right) and CL (left) patients.

Method of fBm estimation is using novel approach and Wishart distribution, that allows to squeeze the whole information from the fBm signal or fBm surface into one matrix. Using the autocorrelation properties, one can obtain the unbiased estimate of Hurst exponent in the whole spectrum, except extreme values, where $H \to 1$. The method of estimation has provided very tiny standard deviation in the estimate and therefore has proven to be suitable for real-data application.

The methodology of Hurst exponent estimate was applied to cranio caudal mammography images, considering the human tissue as a sample of fractional Brownian surface with unknown Hurst exponent. Two approaches involved testing through false discovery rate combined with the Wilcoxon test as well as likelihood ratio test. Both approaches have rejected the null hypothesis that the CL and BL patients have the same Hurst exponent, resulting in patients with cancerous findings having higher Hurst exponent and more chaotic structures in the investigated breast. This is in accordance with the latest research findings in mammogram diagnostics [30].

CRediT authorship contribution statement

Martin Dlask: Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Jaromir Kukal:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the grant No. SGS20/190/OHK4/3T/14 of the Czech Technical University. The author JK acknowledges the support of the OP VVV MEYS, Czech Republic funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”.

References

[1] B.B. Mandelbrot, J.W.V. Ness, Fractional brownian motions, fractional noises and applications, SIAM Rev. 10 (4) (1968) 422–437, http://dx.doi.org/10.1137/1010093.
[2] J. Beran, Statistics for Long-Memory Processes, Chapman & Hall/CRC, Boca Raton London, 1998.
[3] B.B. Mandelbrot, Harold edwin hurst, in: Statisticians of the Centuries, Springer New York, 2001, pp. 335–338, http://dx.doi.org/10.1007/978-1-4613-0179-0_72.
[4] B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman, San Francisco, 1982.

[5] A. Margaris, J.C. Robinson, Embedding properties of sets with finite box-counting dimension, Nonlinearity 32 (10) (2019) 3523–3547, http://dx.doi.org/10.1088/1361-6544/ab1b7f.

[6] M. Dlask, J. Kukal, Application of rotational spectrum for correlation dimension estimation, Chaos, Solitons Fractals 99 (2017) 256–262, http://dx.doi.org/10.1016/j.chaos.2017.04.026.

[7] M. Dlask, J. Kukal, Translation and rotation invariant method of Renyi dimension estimation, Chaos, Solitons Fractals 114 (2018) 536–541, http://dx.doi.org/10.1016/j.chaos.2018.07.030.

[8] P.-O. Amblard, J.-F. Coeurjolly, F. Lavancier, A. Philippe, Basic properties of the multivariate fractional brownian motion, 2012, arXiv:1007.0828.

[9] I. Norros, P. Mannersalo, J. Wang, Simulation of fractional brownian motion with conditionalized random midpoint displacement, Adv. Perform. Anal. 2 (2000).

[10] Z.-M. Yin, New methods for simulation of fractional brownian motion, J. Comput. Phys. 127 (1) (1996) 66–72, http://dx.doi.org/10.1006/jcph.1996.0158.

[11] D.R. McGaughey, G. Aitken, Generating two-dimensional fractional Brownian motion using the fractional Gaussian process (FGp) algorithm, Physica A 311 (3–4) (2002) 369–380, http://dx.doi.org/10.1016/s0378-4371(02)00778-1.

[12] D. Danudirdjo, A. Hirose, Synthesis of two-dimensional fractional Brownian motion via circulant embedding, in: 2011 18th IEEE International Conference on Image Processing, 2011, pp. 1085–1088, http://dx.doi.org/10.1109/ICIP.2011.6115614.

[13] S. Thapa, M.A. Lomholt, J. Krog, A.G. Cherstvy, R. Metzler, Bayesian analysis of single-particle tracking data using the nested-sampling algorithm: maximum-likelihood model selection applied to stochastic-diffusivity data, Phys. Chem. Chem. Phys. 20 (2018) 29018–29037, http://dx.doi.org/10.1039/C8CP04043E.

[14] A.G. Cherstvy, S. Thapa, C.E. Wagner, R. Metzler, Non-Gaussian, non-ergodic, and non-Fickian diffusion of tracers in mucin hydrogels, Soft Matter 15 (2019) 2526–2551, http://dx.doi.org/10.1039/C8SM02096E.

[15] V. Maxim, L. Şendur, J. Fadili, J. Suckling, R. Gould, R. Howard, E. Bullmore, Fractional Gaussian noise, functional MRI and Alzheimers disease, NeuroImage 25 (1) (2005) 141–158, http://dx.doi.org/10.1016/j.neuroimage.2004.10.044.

[16] M. Dlask, J. Kukal, P. Sovka, Fractional brownian bridge model for alzheimer disease detection from EEG signal, in: 2018 International Conference on Signal Processing and Information Security, ICSPIS, IEEE, 2018, http://dx.doi.org/10.1109/cspis.2018.8642720.

[17] O. Lomax, M.L. Bates, A.P. Whitworth, Modelling the structure of star clusters with fractional Brownian motion, Mon. Not. R. Astron. Soc. 480 (1) (2018) 371–380, http://dx.doi.org/10.1093/mnras/sty1788.

[18] I.S. Sahli, Synthesis of mammographic images based on the fractional brownian motion, Int. J. Med. Imaging 6 (1) (2018) 1, http://dx.doi.org/10.11648/j.ijmi.20180601.11.

[19] P.-O. Amblard, J.-F. Coeurjolly, Identification of the multivariate fractional brownian motion, IEEE Trans. Signal Process. 59 (11) (2011) 5152–5168, http://dx.doi.org/10.1109/tsp.2011.2162835.

[20] J. Unterberger, A rough path over multidimensional fractional Brownian motion with arbitrary Hurst index by Fourier normal ordering, Stochastic Process. Appl. 120 (8) (2010) 1444–1472, http://dx.doi.org/10.1016/j.spa.2010.04.001, URL http://www.sciencedirect.com/science/article/pii/S0304414910001067.

[21] J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika 20A (1–2) (1928) 32–52, http://dx.doi.org/10.1093/biomet/20a.1-2.32.

[22] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, I.H. Sloan, Analysis of circulant embedding methods for sampling stationary random fields, SIAM J. Numer. Anal. 56 (3) (2018) 1871–1895, http://dx.doi.org/10.1137/17m1149730.

[23] C.R. Dietrich, G.N. Newsam, Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix, SIAM J. Sci. Comput. 18 (4) (1997) 1088–1107, http://dx.doi.org/10.1137/s1064827592240555.

[24] J. Hoffman, Numerical Methods for Engineers and Scientists, Marcel Dekker, New York, 2001.

[25] R. Gonzalez, Digital Image Processing using MATLAB, Pearson/Prentice Hall, Upper Saddle River, NJ, 2004.

[26] N. Wiener, Time Series, The MIT Press, Cambridge, 1975.

[27] R. Rossi, Mathematical Statistics: An Introduction to Likelihood Based Inference, John Wiley & Sons, Hoboken, NJ, 2018.

[28] M. Heath, K. Bowyer, D. Kopans, R. Moore, P. Kegelmeyer, The digital database for screening mammography, in: Proceedings of the Fourth International Workshop on Digital Mammography, 2000, http://dx.doi.org/10.1007/978-94-011-5318-8_75.

[29] M.D. Heath, K. Bowyer, D. Kopans, W.P. Kegelmeyer, R. Moore, K. Chang, S. Munishkumaran, Current status of the digital database for screening mammography, in: Digital Mammography/IWDM, 1998.

[30] C. Feng, Y. Mei, B. Vidakovic, Mammogram diagnostics using robust wavelet-based estimator of hurst exponent, in: Y. Zhao, D.-G. Chen (Eds.), New Frontiers of Biostatistics and Bioinformatics, Springer International Publishing, Cham, 2018, pp. 109–140, http://dx.doi.org/10.1007/978-3-319-99389-8_5.

PLOS ONE

RESEARCH ARTICLE

Short-time fractal analysis of biological autoluminescence

Martin Dlask[1], Jaromír Kukal[1], Michaela Poplová[2], Pavel Sovka[3], Michal Cifra[2]*

1 Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Trojanova 12, Praha, Czechia, **2** Institute of Photonics and Electronics of the Czech Academy of Sciences, Chaberská 57, Praha 8, Czechia, **3** Department of Circuit Theory of the Faculty of Electrical Engineering at Czech Technical University in Prague, Technická 2, Praha 6, Czechia

* cifra@ufe.cz

OPEN ACCESS

Citation: Dlask M, Kukal J, Poplová M, Sovka P, Cifra M (2019) Short-time fractal analysis of biological autoluminescence. PLoS ONE 14(7): e0214427. https://doi.org/10.1371/journal.pone.0214427

Editor: Miguel Angel Sánchez Granero, Universidad de Almeria, SPAIN

Received: February 27, 2019

Accepted: July 9, 2019

Published: July 26, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: Authors were supported by the following grants: No. SGS17/196/OHK4/3T/14 of the Czech Technical University, CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" of the OP VVV MEYS, internal CTU grant SGS17/183/OHK3/3T/13 "Special Applications of Signal Processing." MC and MP acknowledge Czech Science Foundation project no. 18-23597S.

Abstract

Biological systems manifest continuous weak autoluminescence, which is present even in the absence of external stimuli. Since this autoluminescence arises from internal metabolic and physiological processes, several works suggested that it could carry information in the time series of the detected photon counts. However, there is little experimental work which would show any difference of this signal from random Poisson noise and some works were prone to artifacts due to lacking or improper reference signals. Here we apply rigorous statistical methods and advanced reference signals to test the hypothesis whether time series of autoluminescence from germinating mung beans display any intrinsic correlations. Utilizing the fractional Brownian bridge that employs short samples of time series in the method kernel, we suggest that the detected autoluminescence signal from mung beans is not totally random, but it seems to involve a process with a negative memory. Our results contribute to the development of the rigorous methodology of signal analysis of photonic biosignals.

Introduction

Practically all organisms perpetually generate weak light (300–700 nm wavelength range), too weak to be visible to naked human eye, in the course of their internal metabolic processes [1]. This light phenomenon differs from a rather bright bioluminescence which is dependent on specific enzymatic complexes present only in very specific species such as fireflies and selected jellyfish. What differentiates the general biological autoluminescence from ordinary bioluminescence is, apart from the weaker intensity, its ubiquity across biological species ranging from microorganisms [2–5] through tissue cultures [6–8], plants [9–13] up to animals [14] including human [15–17]. There are also various synonyma used in the literature describing this light phenomenon such as ultra-weak photon emission [18], ultra-weak bioluminescence [19], endogenous biological chemiluminescence [20], biophotons [21–23], etc.

Widely accepted underlying mechanism which generates biological autoluminescence (BAL) is related to a chemical generation of electron-excited states of biomolecules in the course of oxidative metabolism and oxidative stress [18, 24]. While the intensity and optical

Competing interests: The authors have declared that no competing interests exist.

spectrum properties of BAL as a factor of various influences have been widely investigated [3, 25–29], there is limited knowledge and consensus about statistical properties of BAL.

The object of our current study is the BAL time series from the seeds of mung beans that were measured using a sensitive photomultiplier setup. We decided to test the hypothesis if the BAL signals of mung beans contain any intrinsic correlations. To that end, we recorded and analyzed the time series of the BAL from mung beans. One of the ways to assess correlations in the signal employs chaos- and fractal-based approaches [30]. We focus here on the analysis of the fractal character of time series using fractional processes.

Fractional Brownian motion (fBm) and fractional Gaussian noise (fGn), introduced by Mandelbrot [31], have been intensively investigated over the last few decades. They are both dependent on Hurst [32, 33] exponent $H \in (0; 1)$ that influences their autocovariance structure. The fBm or fGn assumption of finite sample is advantageously used in many fields of research of time series analysis—in network traffic modelling [34, 35], financial time series [36, 37], or in biomedicine especially for detection of Alzheimer's disease [38] and cardiology [39].

When analyzing real-world data, the measured sample is usually discrete and short. The traditional methods are generally not suitable for short time series analysis. That is the reason why we need to use a precise method that can estimate the Hurst exponent without bias and can determine the confidence intervals of the estimate. The fractional character of data can be measured via fractional Brownian bridge model, which is a discrete process derived from traditional continuous fBm. A lot of time series are short due to their nature or cut by purpose or experimental limitations. Reconsidering some fBm properties that are taken in long time series analysis as granted and customizing them into a short-time, the discrete model allows estimating Hurst exponent of the discrete measured signal. This approach is advantageously used in a recently developed method of fractional Brownian bridge [40].

The article at first analyzes current open questions of statistical properties of biological autoluminescence. In the next section, we then describe the theory of fBm and the method of Hurst exponent estimation as well as other employed methods, whereas the last section contains the results of the analysis of experimental signals compared to computer-generated reference signals.

Statistical properties of biological autoluminescence (BAL)

Rationale for the need of understanding of BAL statistical properties

Multiple authors proposed that statistical properties of BAL time series might contain an information related to the state of biological system [41–43]. If the existence of such nontrivial statistical properties was rigorously confirmed, it would make a substantial impact on three major areas of this research field.

At first, the discovery of nontrivial statistical properties of BAL would have an impact on the understanding of the BAL generating mechanisms [18, 21]. So far, well-accepted generating mechanism of BAL [18, 24] implicitly considers BAL a weak endogenous biological chemiluminescence formed as a by-product of oxidative metabolism and oxidative stress. General chemiluminescence is typically considered to be random, arising from individual uncorrelated photon emitter molecules [44]. If any correlations in the signal were observed, one would start to ask questions what physical, chemical, and biological processes generate such correlations, hence casting the light on BAL generating mechanisms.

At second, nontrivial statistical properties might revive an interest into long-standing intriguing, yet unresolved question: does BAL enable optical communication between cells and organisms [45–48]? Underlying hypotheses for such biocommunication role of BAL usually expect that BAL carries information which can be processed by a receiver [46]. Such

information could be encoded in the intensity and optical spectrum of BAL [49] or in statistical properties of BAL, if they are any different from random light, as claimed by some authors [45].

At third, statistical properties would represent a completely novel fingerprint for application of BAL in biosensing in biotechnology, agriculture, food industry, and medicine beyond the intensity and optical spectra, hence greatly enhancing application potential of BAL analysis.

Approaches for analysis of BAL statistical properties

Quantum optics approach. Historically, the first common approach to analyze the statistical properties of the photon signals is based on quantum optics theorems and employs photocount statistics of detected photonic signal [50]. Using this approach, several authors suggested that BAL manifests quantum optical coherent properties [21] or even interpreted the observed photocount statistics in terms of quantum optical squeezed states [22, 51]. We have recently criticized the interpretation of experimental evidence claiming quantum optical and quantum coherence properties of BAL [23].

Fractal- and chaos-based signal analysis approach. We believe that it is more realistic to consider that BAL could manifest complex statistical or correlated behavior due to the nature of underlying chemical reactions [52] instead of a hypothetical biological coherent quantum field as proposed in the earlier approach. For the analysis of such complex statistical or correlated behavior, fractal or chaos-based methods seem to be appropriate. Therefore, more recent efforts in the analysis of BAL statistical properties were focused on the various measures quantifying the complexity and correlations in the time series such as Hurst exponent [53] and multifractal spectra [54].

Several works found correlations or deviations from purely random process with a trivial properties in the BAL signal [42, 43, 54]. However, in all those cases, either signals of different signal-to-noise ratio [42, 43, 54] or surrogate (randomly reshuffled time series) [43] were used as reference signals. Comparing BAL signals having different signal-to-noise (signal = net mean intensity of BAL, noise = mean value of detector noise) ratio may lead to results indicating different statistical properties due to a trivial fact: statistical properties of experimentally detected BAL signal are formed by a convolution of detector noise properties with a pure BAL properties. We demonstrated this issue on the example of Fano factor analysis, see figure 4 in [13].

Using surrogate signals might also lead to misleading interpretation in case the signal contains a certain trivial linear trend before random reshuffling—such reshuffling would eradicate any trend. We showed recently that detrending of the BAL signal is not sufficient to remove artifacts since the trend is present not only in the local mean but also in the local variance of the signal (see figures 1b and 4b in [53]). We suggest that the most reliable testing of the hypothesis of nontrivial correlation properties so far can be obtained using reference signals with well-defined properties. To that end, in our recent works, we used computer-generated Poisson signal time series superposed on the experimentally detected detector dark count times series as the control signals with signal-to-noise ratio same as the experimentally detected BAL signals [53]. Such a method for reference signal generation was also recently used in entropy analysis of BAL from model plant *Arabidopsis thaliana* and helped to correctly interpret findings of different entropy values at different stages of seed germination, see figure 6 in [20].

For the first time, we combine here the advanced approach of computer-generated reference signals [53] and a novel method based on fractional Brownian motion analysis [55] to test if BAL signals from mung beans manifest any correlations.

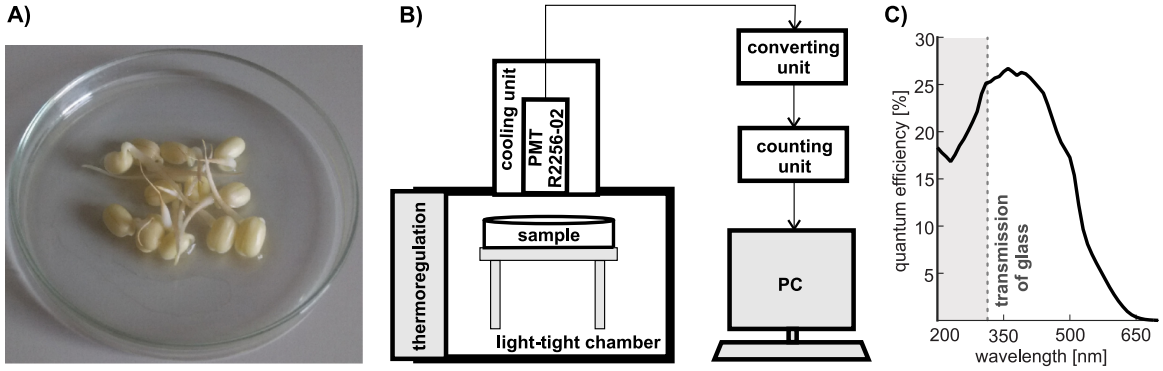


Fig 1. A: Sample of germinating mung beans. B: Scheme of the luminescence measurement setup. C: Quantum efficiency of the photomultiplier used for the detection of biological autoluminescence.

https://doi.org/10.1371/journal.pone.0214427.g001

Materials and methods

Experimental

Preparation of samples. Mung bean seeds (*Vigna radiata*, BIO Mung, CZ-BIO-001) were used as a biological material. Mung seeds were surface-sterilized with 70% ethanol for 1 min. Then, the ethanol was removed, and 50% disinfecting agent (SAVO, CZ) was added. After 10 min, the seeds were washed with distilled water 6 times and soaked for 6 h (shaken every half an hour). After the preparation, the green covers of the seeds were removed. Then, they were germinated in dark condition on large Petri dishes with ultra-pure water.

Luminescence measurement system. We used a measurement system based on cooled (-30 ˚C) low-noise photomultiplier tube (PMT) R2256-02 (all components of the system from Hamamatsu Photonics Deutschland, DE, unless noted otherwise), see Fig 1. Cooling unit C10372 (Hamamatsu Photonics Deutschland, DE) consisted of a control panel and a housing in which the PMT is placed. External water cooling is used for lower cooling temperature. High voltage power supply PS350 (Stanford Research Systems, USA) was used for powering the PMT. C9744 (converting) unit, consisting of a preamplifier, discriminator and a pulse shaping circuit, transforms photocount pulses coming from the PMT into 5V TTL pulses detected by C8855 unit connected to PC. Discriminator level was set to -500 mV and high voltage PMT supply to −1550 V based on the experimental SNR (signal-to-noise-ratio) optimization procedure performed in [56], see figure 2.13 therein.

The PMT had a dark count of ca. 17.2 s^{-1} and photocathode diameter 46 mm); see its quantum efficiency in Fig 1. PMT was mounted from the top outer side of the black light-tight chamber (standard black box, Institute of Photonics and Electronics of the Czech Academy of Sciences, Czechia). The distance between the PMT housing input window and the inner side of the bottom of the Petri dish was 3 cm.

Measurement protocol. The second day after the preparation day, 12 similar mung beans were chosen for the study and distributed into a Petri dish (5 cm in diameter), see Fig 1.

Short sequence analysis

fBm hypothesis. Fractional Brownian motion (fBm) [31] is a continuous Gaussian process $B_H(t)$ defined for continuous variable $t \in [0; +\infty)$, $H \in (0; 1)$ and $\sigma > 0$. The process starts at zero and has zero expected value for all positive times t. The autocovariance structure of

fBm obeys for all $t, s > 0$

$$\mathrm{E}(B_H(t)B_H(s)) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right). \quad (1)$$

Parameter H is called Hurst exponent, for $H = 1/2$, the fBm becomes Wiener process, which is standard Brownian motion. There are several cases of time series behaviour:

- $H \to 1^-$ as strongly dependent and predictable,
- $H \in (1/2; 1)$ as positive long memory process,
- $H = 1/2$ as Wiener-like process,
- $H \in (0; 1/2)$ as negative long memory process,
- $H \to 0^+$ as strongly dependent, but hardly predictable.

Discrete fractional Brownian motion of length $N \in \mathbb{N}$ is any discrete process defined for discrete variable $k = 0, \ldots, N-1$ with zero mean and autocovariance function defined for $k, l = 0, \ldots, N-1$ and $l < N - k$ as

$$\mathrm{E}(B_H(k)B_H(k+l)) = \frac{\sigma^2}{2}\left(|k|^{2H} + |k+l|^{2H} - |l|^{2H}\right). \quad (2)$$

Taking a sample of fractional Brownian motion, it is possible to investigate short samples of time series with fractional character. Finite sample $B_H(k)$ of size $N + 1$ for $k = 0, \ldots, N$ of standardized fBm can be used for the construction of fractional Brownian bridge [55] in the following way

$$M_H(k) = B_H(k) - B_H(0) - \frac{k}{N}(B_H(N) - B_H(0)). \quad (3)$$

In the fractal analysis of time series, the fractional processes are often converted to fractional noises utilizing signal difference to simplify their covariance structure together with its spectral properties keeping the desired dependence on Hurst exponent. The differenced fractional Brownian bridge (dfBB) [55] is defined as

$$X_H(k) = M_H(k+1) - M_H(k) \quad (4)$$

for $k = 0, \ldots, N-1$.

Theory of dfBB. The dfBB is a discrete process and it is proven that the process has zero expected value and its variance is independent on the time lag and equals

$$\gamma_0 = 1 - N^{2H-2}. \quad (5)$$

The autocovariance of dfBB can be expressed as

$$\gamma_m = \eta(m, H) + N^{2H-2} + \frac{|m|^{2H} - |N-m|^{2H} - |N|^{2H}}{N(N-m)}, \quad (6)$$

for $m = 0, 1, \ldots N-1$ where

$$\eta(m, H) = \frac{1}{2}\left(|m+1|^{2H} - 2|m|^{2H} + |m-1|^{2H}\right) \quad (7)$$

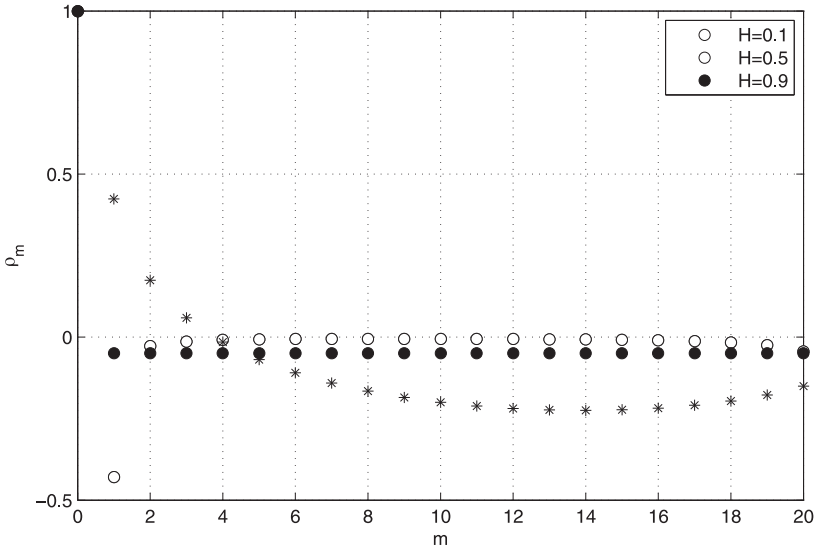


Fig 2. Autocorrelation function of dfBB for $H = 0.1$ (empty circle), $H = 0.5$ (full circle) and $H = 0.9$ (star).

https://doi.org/10.1371/journal.pone.0214427.g002

The corresponding autocorrelation function is again independent on the time lag and can be expressed as

$$\rho_m = \frac{\gamma_m}{\gamma_0} \tag{8}$$

for $m = 0, \ldots, N-1$. The autocorrelation function of dfBB for selected H and $N = 21$ is depicted in Fig 2. This function maps non-negative integer values less than N to the autocorrelation coefficients that will be the key values for subsequent Hurst exponent estimation.

The estimation of Hurst exponent will be based on the correlation function (8). This correlation function is valid only for discrete processes that originated as sampling continuous fBm. In our work, we assume that the investigated signals have the fBm property with unknown Hurst exponent. The advantage of using dfBB is the de-trending of the input signal, which is important in the real experiment outcome analysis.

Hurst exponent estimation. The estimation of Hurst exponent is based on the fitting of the autocorrelation function. For an input discrete signal that has the fBm properties, the dfBB according to formulas (3), (4) is created. If the original signal has length $N + 1$, the respective dfBB has length N having elements $x_0, x_1, \ldots, x_{N-1}$. The estimation of n-th autocovariance coefficient $\hat{r}_n$ can be expressed for $n = 0, \ldots, N-1$ as

$$\hat{r}_n = \frac{1}{N-n} \sum_{k=0}^{N-n-1} x_k x_{k+n} \tag{9}$$

in the case of unbiased estimation. Alternative biased estimate is based on formula

$$\hat{r}_n = \frac{1}{N} \sum_{k=0}^{N-n-1} x_k x_{k+n} \tag{10}$$

and the estimation of autocorrelation coefficient $\hat{\rho}_n$ as

$$\hat{\rho}_n = \frac{\hat{r}_n}{\hat{r}_0}. \tag{11}$$

The results using (9) and (10) were proven to be comparable, therefore we used the Eq (9) for the following calculations. Denote the theoretical value of autocorrelation from Eq (8) as $\rho_n = \rho_n(H)$ and the experimentally calculated autocorrelation as $\hat{\rho}_n$. Then we obtain the estimation of parameter H by means of solving the minimization problem

$$\hat{H} = \underset{H\in(0,1)}{\operatorname{argmin}} \sum_{j=1}^{M}\sum_{n=1}^{N-1}(\rho_{n,j} - \rho_n(H))^2, \tag{12}$$

where M is the number of signal segments. The point estimate of $\hat{H}$ was obtained by the maximum likelihood method [57] together with its standard deviation $\hat{s}$ as recommended in [55].

Likelihood ratio test

Having signal from the mung beans photon emission as well as the reference signal, we will use likelihood ratio test [58] to decide, whether the Hurst exponent of both samples is significantly different. We denote H_D as the Hurst exponent estimate of the PMT detector noise or reference signal and H_B as the Hurst exponent estimate of mung emission using the formula (12). The overall error (sum of the squares of residuals) is defined as

$$SSQ_{FULL} = \sum_{i=1}^{M}\sum_{j=1}^{N-1}(\rho_{i,j}^{B} - \rho_j(H_B))^2 + \sum_{i=1}^{M}\sum_{j=1}^{N-1}(\rho_{i,j}^{D} - \rho_j(H_D))^2, \tag{13}$$

where ρ^D, ρ^B are the autocorrelation coefficient of the noise and photon emission, respectively. The case of $j = 0$ is excluded due to $\rho_{i,0}^{D} = \rho_0(H_D) = 1$ for all $i = 1, .., M$. Using sub-model satisfying $H_B = H_D$ we get

$$SSQ_{SUB} = \sum_{i=1}^{M}\sum_{j=1}^{N-1}(\rho_{i,j}^{B} - \rho_j(H_D))^2 + \sum_{i=1}^{M}\sum_{j=1}^{N-1}(\rho_{i,j}^{D} - \rho_j(H_D))^2. \tag{14}$$

Using likelihood ratio (LR) test of significant difference between the sub-model and the full model, we calculate

$$\chi^2 = 2\ln\frac{L_{FULL}}{L_{SUB}} = M\cdot(N-1)\cdot\ln\frac{SSQ_{SUB}}{SSQ_{FULL}}, \tag{15}$$

where L_{FULL} and L_{SUB} are corresponding likelihoods.

When the hypothesis H_0: $H_D = H_B$ holds, i.e. the full model has the same validity as the sub-model, the criterion has χ_1^2 distribution due to single parameter constrain.

Reference signal generation. Recently we demonstrated that a suitable reference signal is crucial to understand and interpret the findings from various BAL signal analysis [20, 53]. Detector noise itself is not a suitable reference signal since it contains intrinsic technogenic correlations itself [53] and using signals of other samples with different signal-to-detector noise ratio can also lead to misleading results as we explained in section "Approaches for analysis of BAL statistical properties". Hence for this work we follow our method [53], and generated the reference signal as a sum of measured detector noise and computer-generated Poisson signal (using Matlab®2017 *poissrnd* command) with given λ in every experimental point

Table 1. Mean values of mung beans signal and noise.

T_b	200 μs	500 μs
$\mathrm{E}\, y_k^{\mathrm{B}}$	0.0115	0.0288
$\mathrm{E}\, y_k^{\mathrm{D}}$	0.0036	0.0088
λ	0.0079	0.0200

https://doi.org/10.1371/journal.pone.0214427.t001

where $\lambda = \mathrm{E}\, y_k^{\mathrm{B}} - \mathrm{E}\, y_k^{\mathrm{D}}$, where y_k^{B} and y_k^{D} are signal mean values of mung beans and noise, respectively. The respective values of λ in case of 200 μs signal as well as 500 μs signal are calculated in Table 1. Hence, experimentally detected BAL signals from mung beans and reference signals have practically the same mean value and same signal-to-noise ratio.

To sum up, for the analysis, we have three types of signals available:

- (B)—mung beans signal y_k^{B},
- (D)—noise signal of PMT detector y_k^{D},
- (R)—reference signal as a sum of measured detector noise (D) and computer-generated Poisson noise denoted as y_k^{R}

Results

Measurement

The investigated sample of germinating mung beans is displayed in the Fig 1. An overview of all signals collected and employed in this paper is in Table 2, where NS denotes the number of available signals.

There were two bin size settings used to collect the signals: T_s = 200 and 500 μs. For each sampling period, we have corresponding mung bean signals, detector noise signals, and computer-generated reference signals.

Both mung beans signal and PMT detector noise signal are assumed to be stationary with their mean values with the Poisson distribution. Therefore, they can be represented by their mean values $\mathrm{E}\, y_k^{\mathrm{B}}$ and $\mathrm{E}\, y_k^{\mathrm{D}}$ that are estimated from the measured data.

As previously mentioned, the aim of study is to compare mung beans signal with the reference signal and find statistical difference between them using their autocorrelation. With each of these two signals independently, we performed basic data processing. This procedure describes the normalization of the data, which is the essential property of fBm processes. At

Table 2. Number and type of the signals collected.

bin size T_b		200 μs	500 μs
signal type	mung beans (B)	NS = 5	NS = 5
	detector noise (D)	NS = 5	NS = 5
	reference signals (R)	NS = 5	NS = 5
number of bins in each measurement		N_b = 100 000	N_b = 100 000
length of each measurement [s]		20	50
total number of bins Q = NS × N_b		500 000	500 000
total length of all measurements per signal type [s]		100	500

https://doi.org/10.1371/journal.pone.0214427.t002

first the input time series y_k for $k = 0, 1, \ldots, Q-1$ was cumulatively summed for a window size $h \in \mathbb{N}$ and Anscombe transformation [59] was performed. The resulting signal z_k can be expressed based on the output from measuring device y_k as

$$z_k = 2 \cdot \left(\frac{3}{8} + \sum_{i=kh}^{(k+1)h-1} y_i \right)^{1/2} \tag{16}$$

for $k = 0, \ldots, M-1$. This transformations assures stationarity by terms of variance and guarantees Gaussian distribution of the resulting signal.

Hurst exponent estimates

There is no prior knowledge of optimal model length, accumulation compression, and Hurst exponent. Therefore, we will apply the maximum likelihood method of Hurst exponent estimation for the various model and segment lengths, and then we will individually test the differences in the Hurst exponent. However, there is a finite number of reasonable pairs (model length N, segment length h), which will cause the phenomenon of the multiple hypothesis testing. After the False Discovery Rate (FDR) correction, we will localize the model and segment lengths, which cause significant differences in the Hurst exponent. These pairs (h, N) will be declared as significantly sensitive to the signal differences in the Hurst exponent.

Having signals with two different bin sizes, we will use the signal bin size $T_b = 200$ μs as a training set and the signal with $T_b = 500$ μs as a verification set. Normalized mung beans and reference signals with bin size $T_b = 200$ μs and length $Q = 500\,000$ were the subject of the initial analysis. The signal accumulation of size h was applied to the signals, therefore the number of bins was $\lfloor Q/h \rfloor$. After the accumulation, the signal is divided into segments of length N. Due to the memory of fBm process, we will use only the odd segments for the calculation of autocorrelation function and the even segments are excluded. The new signal has length $\lceil \lfloor \lfloor Q/h \rfloor / N \rfloor / 2 \rceil$. Using Eq (12) and maximum likelihood method, we obtain the corresponding H_D and H_B estimates for the Hurst exponent of referential signal and mung beans, respectively. Based on these estimates, we can derive the p-values of LR test using (15) statistics.

In our case, we performed altogether $11 \times 11 = 121$ tests for $h = 1500, 1550, \ldots, 2000$ and $N = 20, 21, \ldots, 30$. Accumulation h could not be higher than 2000 due to the rapid decrease of the number of processed segments. The values $h < 1500$ caused lower event frequencies and the conversion from Poisson noise to Gaussian noise is not guaranteed. Similar reasons are for the range of parameter N. In fact, the fractional model is less discriminative for $N < 20$ and the case $N > 30$ reduces the number of segments. Due to multiple testing and obeying the Hochberg-Benjamini principle, we diminish the significance level from 0.05 to $\alpha_{FDR} = 0.000050$. The p-values as decadic logarithms are shown in Table 3.

In these settings, there were two cases where the Hurst exponent was significantly different. The results from these two cases are displayed in Table 4. The 95% interval of Hurst exponent estimates of mungo beans signal (H_B) was [0.2108, 0.5086], while the 95% interval for reference signal (H_R) was [0.4041,0.5931].

The lowest p-value was obtained in the case of $(h, N) = (1750, 24)$, which represents the segmentation into bins with duration 1750×200 μs = 350 000 μs = 0.35 sec.

As the verification set, the signal with $T_s = 500$ μs was taken into account, following the same procedure as the previous one. The accumulation parameter h was accordingly diminished to 2/5 of its previous value to guarantee the same segment length.

Table 3. Difference between the estimated Hurst exponent of mung beans (B) and reference signal (R) as ($-\log_{10} p$)-values of likelihood ratio test (15).

h \N	20	21	22	23	24	25	26	27	28	29	30
1500	1.188	2.934	1.835	1.888	3.175	1.645	2.284	0.863	0.506	0.192	0.762
1550	1.172	1.617	0.394	0.887	1.420	1.470	0.912	2.113	1.651	0.026	0.691
1600	1.978	1.576	0.646	0.523	1.217	0.394	1.597	0.786	1.487	0.880	1.859
1650	0.990	1.616	1.127	2.024	1.209	0.651	1.635	0.909	1.906	3.573	2.927
1700	0.772	0.621	1.288	1.196	1.239	0.488	0.407	1.175	2.658	0.463	0.776
1750	1.475	2.325	1.269	3.131	**4.638**	1.535	2.370	1.017	0.726	0.412	1.945
1800	0.465	1.455	1.394	1.098	1.313	0.180	2.661	2.064	2.449	1.917	2.001
1850	2.377	2.010	1.308	0.567	1.533	2.382	3.184	**4.301**	3.328	2.418	1.968
1900	2.599	0.879	0.850	0.629	1.053	1.264	0.950	0.943	1.397	2.093	0.142
1950	2.574	0.095	0.706	1.900	2.843	2.874	3.261	2.514	3.462	2.501	2.405
2000	2.212	1.611	1.315	0.935	1.040	1.232	0.922	0.282	0.366	1.159	0.963

https://doi.org/10.1371/journal.pone.0214427.t003

Table 4. Estimated Hurst exponent values for mung beans (B) signal and reference signal (R).

h	N	H_B	H_R	p-val	$-\log_{10} p$-val
1750	24	0.4142	0.5299	2.3135×10^{-5}	4.638
1850	27	0.3569	0.4291	4.9977×10^{-5}	4.301

https://doi.org/10.1371/journal.pone.0214427.t004

We perform the verification for the combination of signals (B) and (R) similarly as in the previous case and additionally for the combination of (B) and (D). The first set of signals ((B) and (R)) will be used to test if the photon emission is not random and has a negative memory, while the results from the second set ((B) and (D)) of signals will be used to test if there is a significant difference between the cases, when the PMT detects BAL signals from mung beans compared to PMT noise. We use the significant cases from Table 4 to estimate their Hurst exponent and the results on verification set is displayed in Table 5. The variables s_1, s_2 denote the pair of signals, whereas the H_X denotes the estimation of Hurst exponent of the signal s_2.

We performed four tests, and according to Hochberg-Benjamini false discovery rate, we diminish the $\alpha_{FDR} = 0.0169$. Therefore, all four cases are considered significant, and we reject the hypothesis that the Hurst exponent of mung beans would be the same as H_X.

For comparison, we also performed a similar analysis with noise signal (D) and mungo beans signal (B) and captured the results in the Table 6. Using Hochberg-Benjamini principle, there is only one combination (h,N) = (1850,27) that is significant.

To demonstrate the efficiency of proposed method for short time series, we applied the method of power spectral analysis (PSD) [60, 61] for Hurst exponent estimation. The power spectrum $P(f)$ holds following relationship for any frequency f

$$P(f) \propto f^{-1-2H}, \tag{17}$$

Table 5. Estimated Hurst exponent values from verification dataset. h = 700 for 500 μs signals corresponds to h = 1750 for 200 μs signals.

s_1	s_2	h	N	H_B	H_X	p-val
B	R	700	24	0.4032	0.4415	0.0130
B	R	740	27	0.3761	0.4112	0.0042
B	D	700	24	0.4032	0.4378	0.0169
B	D	740	27	0.3761	0.4480	0.0054

https://doi.org/10.1371/journal.pone.0214427.t005

Table 6. Difference between the estimated Hurst exponent of mung beans (B) and noise signal (D) as ($-\log_{10} p$)-values of likelihood ratio test (15).

h \N	20	21	22	23	24	25	26	27	28	29	30
1500	2.181	2.881	1.395	1.656	3.192	2.058	2.203	0.087	0.444	1.131	3.319
1550	2.191	1.125	0.446	0.736	0.348	2.101	0.283	2.378	1.936	1.560	0.407
1600	2.008	2.039	1.259	0.006	1.071	1.098	1.604	2.003	0.665	1.468	1.415
1650	1.230	1.535	1.334	2.804	1.249	0.914	0.858	1.470	2.923	2.190	3.056
1700	1.015	1.359	0.376	0.460	0.473	0.665	0.182	0.906	1.812	1.532	0.107
1750	0.769	1.218	0.926	0.346	1.771	2.386	0.687	0.331	1.286	1.136	1.605
1800	0.696	1.791	0.871	0.666	2.058	2.153	2.053	1.145	2.400	0.692	0.881
1850	2.891	1.967	1.349	0.297	1.051	2.343	2.754	**4.602**	3.472	0.779	1.432
1900	2.795	0.434	0.708	1.273	0.619	0.631	1.797	1.101	0.395	1.371	1.069
1950	1.657	1.694	0.386	2.885	2.373	2.199	2.504	2.292	1.651	0.926	2.209
2000	1.281	0.048	0.215	1.050	0.140	1.031	0.234	1.303	1.186	0.239	0.939

https://doi.org/10.1371/journal.pone.0214427.t006

therefore the Hurst exponent can be estimated using the maximum likelihood method in log-log chart. The results from the spectral method from training data are captured in Table 7, together with the Hurst exponent estimates of referential and mungo beans signals. The method exhibited similar results, showing that the estimates of H_B are lower than H_D; however, due to the very short signal length, the method provided huge standard deviation of the estimate. For this reason, the p-values are not significant, and all of them are above the value of 0.5. The obtained results were similar in the case of verification data as well, showing no significant records in the whole range of parameter *h*.

To assure the both PSD and differenced fractional Brownian bridge (DFBB) methods are unbiased, we generated artificial signal sample of length 300 using circular embedding method [62], which is an exact method for fractional Brownian motion generation. The results of this testing are captured in Table 8. While *H* denotes the Hurst exponent of artificially generated

Table 7. Spectral analysis of mung beans and reference signal as *p*-values.

h	H_B	H_R	*p*-val
1500	0.4366	0.4425	0.9039
1550	0.4871	0.5269	0.8210
1600	0.4661	0.4592	0.9436
1650	0.4107	0.4358	0.9202
1700	0.4551	0.4502	0.9642
1750	0.4322	0.5591	0.5625
1800	0.3827	0.4173	0.6539
1850	0.4366	0.4406	0.9043
1900	0.4257	0.4312	0.8789
1950	0.3973	0.4046	0.8462
2000	0.4715	0.6516	0.6821

https://doi.org/10.1371/journal.pone.0214427.t007

Table 8. Unbiasedness of PSD and DFBB methods as p-value.

H	H^{PSD}	s^{PSD}	p-value	H^{DFBB}	s^{DFBB}	p-value
0.35	0.4268	0.1956	0.3471	0.3513	0.0193	0.4728
0.40	0.4152	0.2211	0.4725	0.4123	0.0205	0.2741
0.45	0.4803	0.1676	0.4786	0.4335	0.0249	0.7443
0.50	0.5454	0.2198	0.4840	0.5030	0.0276	0.4566

https://doi.org/10.1371/journal.pone.0214427.t008

signal, the H^{PSD}, H^{DFBB} denote the Hurst exponent estimates and s^{PSD}, s^{DFBB} denote the standard deviation of this estimate using the PSD and DFBB method, while the p-value denotes the result of the t-test against the null hypothesis that the estimated value equals the theoretical one. Having all p-values above 0.05 significance level, we conclude that both PSD and DFBB method provide unbiased estimates.

Discussion

Results from rigorous statistical analysis and testing in Tables 3, 4, and 5 suggest that the mung beans signal has a negative memory (negative correlations, antipersistent behavior [63]) and its Hurst exponent is lower than the referential signal. How could such behavior originate in biological systems? It was proposed that the restriction of Brownian motion due to the structuring of nano- to microscale intracellular environment leads to anomalous sub-diffusion [64] characterized by Hurst exponent < 0.5 [63]. This is understandable since a cytoplasm environment displays fractal spatial structuring [65]. Since biochemical reactions (encounters of reactants) leading to BAL are taking place within the cell cytoplasm, organelles and lipid membranes [24] where anomalous sub-diffusion was observed [64, 66], it is not a great logical leap to speculate that BAL from mung bean samples could also display sub-diffusive features. Actually, it is already acknowledged that chemical reactions spatially constrained on the microscopic level may lead to fractal reaction kinetics [67–69] also in case of intracellular biochemical kinetics [70]. The 0.35 s as the time scale where we found statistically significant differences of mung bean signal Hurst exponent from that of the reference signal (Table 3) could correspond to a rate of underlying rate-limiting step of chemical reactions or processes which give rise to BAL. However, one has to be careful in the interpretation since there are many pitfalls in an accurate estimation of the Hurst exponent value from experiments [71, 72]. Although unlikely, given the nature of our experiments we can not fully exclude that the correlations we observe in mung signals are introduced by the photodetector (PMT) due to the nature of photocounting process [73, 74]. Introduction of anti-/correlations could be at the physical level of the PMT tube (after-pulsing, a temporary drop of the voltage at dynodes after ejecting electrons, . . .) or the follow-up circuitry (amplifiers). Anti-correlations of the detected counts depending on the count rate have been actually observed due to a PMT construction [74, Fig.9]. However, marked anti-correlations were present only for very high count rates ($>$ kHz) and very low quantum efficiency, which is not the case in our experiments. We also believe that the dead-time of a PMT [75] is not affecting the value of correlations we observe since the PMT dead-time is on the time scale of few hundreds of nanoseconds—several orders of magnitude smaller than the time scale of correlations we observed (0.35 s) and three orders of magnitude smaller than our bin size (200 and 500 μs).

Throughout the analysis, the lower limit for accumulation parameter h was chosen as 1500 to assure the normality of the processed data due to the sparsity of the input signal. Higher accumulation than 2000 is not useful since then we would lose the precision of estimate due to the short length of investigated time series. The minimal length of signal segment N was chosen to assure consistency of the used model, segment lengths of $N > 30$ do not significantly contribute to the higher precision of estimate [40].

Conclusion

In this work, we focused on statistical properties of biological autoluminescence from germinating mung bean sample. Our emphasis was on the development of a rigorous mathematical

and statistical methodology, which takes into account proper reference signals, likelihood ratio test, and multiple hypothesis testing effects.

We used a highly sensitive photomultiplier-based detection system to record a time series of photon counts of the mung bean sample emission and noise of the detector. Using the normalization of the input signal, we were able to employ the fractional models that allowed us to estimate Hurst exponent. Dividing the input signals into the training set and evaluating the differences in the Hurst exponent of both signals, the procedure allowed us to test our initial hypothesis on the verification signal. The resulting Hurst exponent mean value of mung bean sample time series is below the level of 1/2 which confirmed our initial hypothesis, that the biological autoluminescence displays correlations. We also proposed that this value could be related to anomalous sub-diffusive features of biochemical reactions underlying processes within mung beans, which give rise to photon emission time series. Further extensive work beyond the scope of this methodical paper needs to be carried out to test the biological ubiquity of anti-/correlations in biological autoluminescence signals and the role of the detector in the observed Hurst exponent values. Especially interesting would be an analysis of BAL statistical properties across samples with rising complexity starting from simple chemical solutions of small biomolecules through isolated cellular structures and cell suspensions up to whole tissues and organisms. Nevertheless, we believe that rigorous methodology we presented here will help to support the future research of BAL statistical properties towards a deeper understanding of BAL mechanisms as well as applications for label-free and non-invasive analysis in medicine and biotechnology using completely new signal fingerprint types.

Supporting information

S1 Data. All raw data files used in our analysis.
(ZIP)

Author Contributions

Conceptualization: Michal Cifra.

Data curation: Martin Dlask, Michal Cifra.

Formal analysis: Martin Dlask.

Funding acquisition: Michal Cifra.

Investigation: Martin Dlask, Michaela Poplová.

Methodology: Jaromír Kukal, Pavel Sovka, Michal Cifra.

Project administration: Jaromír Kukal, Michal Cifra.

Resources: Jaromír Kukal, Michal Cifra.

Software: Martin Dlask, Michaela Poplová.

Supervision: Jaromír Kukal, Pavel Sovka, Michal Cifra.

Validation: Michal Cifra.

Visualization: Martin Dlask, Michaela Poplová.

Writing – original draft: Martin Dlask, Michal Cifra.

Writing – review & editing: Jaromír Kukal, Pavel Sovka, Michal Cifra.

References

1. Kucera O, Cervinkova K, Nerudova M, Cifra M. Spectral perspective on the electromagnetic activity of cells. Current topics in medicinal chemistry. 2015; 15(6):513–522. https://doi.org/10.2174/1568026615666150225103105 PMID: 25714382
2. Tilbury R, Quickenden T. Luminescence from the yeast Candida utilis and comparisons across three genera. Luminescence. 1992; 7(4):245–253.
3. Quickenden TI, Tilbury RN. Luminescence spectra of exponential and stationary phase cultures of respiratory deficient Saccharomyces cerevisiae. Journal of Photochemistry and Photobiology B: Biology. 1991; 8(2):169–174. https://doi.org/10.1016/1011-1344(91)80055-M
4. Vogel R, Guo X, Süssmuth R. Chemiluminescence patterns from bacterial cultures undergoing bacteriophage induced mass lysis. Bioelectrochemistry and Bioenergetics. 1998; 46(1):59–64.
5. Vogel R, Süssmuth R. Weak light emission patterns from lactic acid bacteria. Luminescence. 1999; 14 (2):99–105. https://doi.org/10.1002/(SICI)1522-7243(199903/04)14:2<99::AID-BIO519>3.0.CO;2-7 PMID: 10398567
6. Burgos RCR, Červinková K, van der Laan T, Ramautar R, van Wijk EPA, Cifra M, et al. Tracking biochemical changes correlated with ultra-weak photon emission using metabolomics. Journal of Photochemistry and Photobiology B: Biology. 2016; 163:237–245. https://doi.org/10.1016/j.jphotobiol.2016.08.030
7. Burgos RCR, Schoeman JC, Winden LJv, Červinková K, Ramautar R, Van Wijk EPA, et al. Ultra-weak photon emission as a dynamic tool for monitoring oxidative stress metabolism. Scientific Reports. 2017; 7(1):1229. https://doi.org/10.1038/s41598-017-01229-x PMID: 28450732
8. Rác M, Sedlářová M, Pospíšil P. The formation of electronically excited species in the human multiple myeloma cell suspension. Scientific Reports. 2015; 5:8882. https://doi.org/10.1038/srep08882 PMID: 25744165
9. de Mello Gallep C. Ultraweak, spontaneous photon emission in seedlings: toxicological and chronobiological applications: UPE in seedlings—applications. Luminescence. 2014; 29(8):963–968. https://doi.org/10.1002/bio.2658 PMID: 24687546
10. Gallep CM, Moraes TA, Dos Santos SR, Barlow PW. Coincidence of biophoton emission by wheat seedlings during simultaneous, transcontinental germination tests. Protoplasma. 2013; 250(3):793–796. https://doi.org/10.1007/s00709-012-0447-x PMID: 23011402
11. Gallep CM, Moraes TA, Červinková K, Cifra M, Katsumata M, Barlow PW. Lunisolar tidal synchronism with biophoton emission during intercontinental wheat-seedling germination tests. Plant Signaling & Behavior. 2014; 9(5):e28671. https://doi.org/10.4161/psb.28671
12. Guo J, Zhu G, Li L, Liu H, Liang S. Ultraweak photon emission in strawberry fruit during ripening and aging is related to energy level. Open Life Sciences. 2017; 12(1). https://doi.org/10.1515/biol-2017-0046
13. Rafieiolhosseini N, Poplová M, Sasanpour P, Rafii-Tabar H, Alhossaini MR, Cifra M. Photocount statistics of ultra-weak photon emission from germinating mung bean. Journal of Photochemistry and Photobiology B: Biology. 2016; 162:50–55. https://doi.org/10.1016/j.jphotobiol.2016.06.001
14. van Wijk E, Kobayashi M, van Wijk R, van der Greef J. Imaging of Ultra-Weak Photon Emission in a Rheumatoid Arthritis Mouse Model. PLoS ONE. 2013; 8(12):e84579. https://doi.org/10.1371/journal.pone.0084579 PMID: 24386396
15. Ou-Yang H. The application of ultra-weak photon emission in dermatology. Journal of Photochemistry and Photobiology B: Biology. 2014; 139:63–70. https://doi.org/10.1016/j.jphotobiol.2013.10.003
16. Zhao X, Wijk Ev, Yan Y, Wijk Rv, Yang H, Zhang Y, et al. Ultra-weak photon emission of hands in aging prediction. Journal of Photochemistry and Photobiology B: Biology. 2016; 162:529–534. https://doi.org/10.1016/j.jphotobiol.2016.07.030
17. Zhao X, Pang J, Fu J, Wang Y, Yang M, Liu Y, et al. Spontaneous photon emission: A promising non-invasive diagnostic tool for breast cancer. Journal of Photochemistry and Photobiology B: Biology. 2017; 166:232–238. https://doi.org/10.1016/j.jphotobiol.2016.12.009
18. Cifra M, Pospíšil P. Ultra-weak photon emission from biological samples: Definition, mechanisms, properties, detection and applications. Journal of Photochemistry and Photobiology B: Biology. 2014; 139:2–10. https://doi.org/10.1016/j.jphotobiol.2014.02.009
19. Wang J, Yu Y. Relationship between ultra-weak bioluminescence and vigour or irradiation dose of irradiated wheat. Luminescence. 2009; 24(4):209–212. https://doi.org/10.1002/bio.1096 PMID: 19291806
20. Saeidfirozeh H, Shafiekhani A, Cifra M, Masoudi AA. Endogenous Chemiluminescence from Germinating Arabidopsis Thaliana Seeds. Scientific Reports. 2018; 8(1):16231. https://doi.org/10.1038/s41598-018-34485-6 PMID: 30385859

21. Popp FA, Nagl W, Li K, Scholz W, Weingärtner O, Wolf R. Biophoton emission: New evidence for coherence and DNA as source. Cell Biochemistry and Biophysics. 1984; 6(1):33–52. https://doi.org/10.1007/BF02788579

22. Bajpai RP. Biophoton emission in a squeezed state from a sample of Parmelia tinctorum. Physics Letters A. 2004; 322(1-2):131–136. https://doi.org/10.1016/j.physleta.2003.12.050

23. Cifra M, Brouder C, Nerudová M, Kučera O. Biophotons, coherence and photocount statistics: A critical review. Journal of Luminescence. 2015; 164:38–51. https://doi.org/10.1016/j.jlumin.2015.03.020

24. Pospíšil P, Prasad A, Rác M. Role of reactive oxygen species in ultra-weak photon emission in biological systems. Journal of Photochemistry and Photobiology B: Biology. 2014; 139:11–23. https://doi.org/10.1016/j.jphotobiol.2014.02.008

25. Prasad A, Pospísil P. Ultraweak photon emission induced by visible light and ultraviolet A radiation via photoactivated skin chromophores: in vivo charge coupled device imaging. Journal of biomedical optics. 2012; 17(8):085004. https://doi.org/10.1117/1.JBO.17.8.085004 PMID: 23224187

26. Yang M, Pang J, Liu J, Liu Y, Fan H, Han J. Spectral discrimination between healthy people and cold patients using spontaneous photon emission. Biomedical Optics Express. 2015; 6(4):1331. https://doi.org/10.1364/BOE.6.001331 PMID: 25909016

27. Kobayashi M, Iwasa T, Tada M. Polychromatic spectral pattern analysis of ultra-weak photon emissions from a human body. Journal of Photochemistry and Photobiology B: Biology. 2016; 159:186–190. https://doi.org/10.1016/j.jphotobiol.2016.03.037

28. Nerudová M, Červinková K, Hašek J, Cifra M. Optical spectral analysis of ultra-weak photon emission from tissue culture and yeast cells. In: Tománek P, Senderáková D, Páta P, editors. SPIE Proceedings; 2015. p. 94500O. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2069897.

29. Zhao X, Yang M, Wang Y, Pang J, Wijk EV, Liu Y, et al. Spectrum of spontaneous photon emission as a promising biophysical indicator for breast cancer research. Scientific Reports. 2017; 7(1).

30. Eke A, Herman P, Kocsis L, Kozak LR. Fractal characterization of complexity in temporal physiological signals. Physiological Measurement. 2002; 23(1):R1–R38. https://doi.org/10.1088/0967-3334/23/1/201 PMID: 11876246

31. Mandelbrot B, Ness JV. Fractional Brownian Motions, Fractional Noises and Applications. SIAM Review. 1968; 10(4):422–437. https://doi.org/10.1137/1010093

32. Hurst HE. Methods of Using Long-Term Storage in Reservoirs. Proceedings of the Institution of Civil Engineers. 1956; 5(5):519–543. https://doi.org/10.1680/iicep.1956.11503

33. Julian M, Alcaraz R, Rieta JJ. Application of Hurst exponents to assess atrial reverse remodeling in paroxysmal atrial fibrillation. Physiol Meas. 2015; 36(11):2231–2246. https://doi.org/10.1088/0967-3334/36/11/2231 PMID: 26393825

34. Erramilli A, Willinger W, Pruthi P. Fractal traffic flows in high-speed communications networks. Fractals. 1994; 2(03):409–412. https://doi.org/10.1142/S0218348X94000545

35. Li D, Fan Q. Multi-fractal Modeling of Network Video Traffic and Performance Analysis. Journal of Internet Technology. 2018; 19(7):2089–2095.

36. Dubovikov MM, Starchenko NV, Dubovikov MS. Dimension of the minimal cover and fractal analysis of time series. Physica A: Statistical Mechanics and its Applications. 2004; 339(3-4):591–608. https://doi.org/10.1016/j.physa.2004.03.025

37. Bohdalová M, Greguš M. Fractal Analysis of Forward Exchange Rates. Acta Polytechnica Hungarica. 2010; 7(4):14.

38. Dlask M, Kukal J, Sovka P. Fractional Brownian Bridge Model for Alzheimer Disease Detection from EEG Signal. In: 2018 International Conference on Signal Processing and Information Security (ICSPIS); 2018. p. 1–4.

39. Captur G, Karperien AL, Hughes AD, Francis DP, Moon JC. The fractal heart—embracing mathematics in the cardiology clinic. Nature Reviews Cardiology. 2017; 14(1):56–64. https://doi.org/10.1038/nrcardio.2016.161 PMID: 27708281

40. Dlask M, Kukal J. Hurst exponent estimation from short time series. Signal, Image and Video Processing. 2018;.

41. Kobayashi M, Inaba H. Photon statistics and correlation analysis of ultraweak light originating from living organisms for extraction of biological information. Applied Optics. 2000; 39(1):183–192. https://doi.org/10.1364/ao.39.000183 PMID: 18337887

42. Van Wijk R, Van Wijk EPA, Bajpai RP. Photocount distribution of photons emitted from three sites of a human body. Journal of Photochemistry and Photobiology B: Biology. 2006; 84(1):46–55. https://doi.org/10.1016/j.jphotobiol.2006.01.010

43. Van Wijk EPA, Wijk RV, Bajpai RP, van der Greef J. Statistical analysis of the spontaneously emitted photon signals from palm and dorsal sides of both hands in human subjects. Journal of Photochemistry and Photobiology B: Biology. 2010; 99(3):133–143. https://doi.org/10.1016/j.jphotobiol.2010.03.008

44. Collinson MM, Wightman RM. Observation of individual chemical reactions in solution. Science. 1995; 268(5219):1883. https://doi.org/10.1126/science.268.5219.1883 PMID: 17797530

45. Budagovsky AV. On the ability of cells to distinguish the coherence of optical radiation. Quantum Electronics. 2005; 35(4):369–374. https://doi.org/10.1070/QE2005v035n04ABEH002837

46. Kučera O, Cifra M. Cell-to-cell signaling through light: just a ghost of chance? Cell Communication and Signaling. 2013; 11(1):1.

47. Prasad A, Rossi C, Lamponi S, Pospíšil P, Foletti A. New perspective in cell communication: Potential role of ultra-weak photon emission. Journal of Photochemistry and Photobiology B: Biology. 2014; 139:47–53. https://doi.org/10.1016/j.jphotobiol.2014.03.004

48. Scholkmann F, Fels D, Cifra M. Non-chemical and non-contact cell-to-cell communication: a short review. American journal of translational research. 2013; 5(6):586. PMID: 24093056

49. Laager F. Light based cellular interactions: hypotheses and perspectives. Frontiers in Physics. 2015; 3. https://doi.org/10.3389/fphy.2015.00055

50. Walls DF, Milburn GJ. Quantum optics. Springer; 2008.

51. Bajpai RP. Squeezed state description of spectral decompositions of a biophoton signal. Physics Letters A. 2005; 337(4-6):265–273. https://doi.org/10.1016/j.physleta.2005.01.079

52. Iranifam M, Segundo MA, Santos JLM, Lima JLFC, Sorouraddin MH. Oscillating chemiluminescence systems: state of the art. Luminescence. 2010; 25(6):409–418. https://doi.org/10.1002/bio.1203 PMID: 20354969

53. Poplová M, Sovka P, Cifra M. Poisson pre-processing of nonstationary photonic signals: Signals with equality between mean and variance. PLOS ONE. 2017; 12(12):e0188622. https://doi.org/10.1371/journal.pone.0188622 PMID: 29216207

54. Scholkmann F, Cifra M, Moraes TA, de Mello Gallep C. Using multifractal analysis of ultra-weak photon emission from germinating wheat seedlings to differentiate between two grades of intoxication with potassium dichromate. Journal of Physics: Conference Series. 2011; 329:012020.

55. Dlask M, Kukal J, Tran QV. Revisited Zero-Crossing Method for Hurst Exponent Estimation in Time Series. In: Martincik D, editor. Mathematical Methods in Economics Proceedings 2015. vol. 1. University of West Bohemia, Cheb, Czech Republic: University of West Bohemia, Plzen; 2015. p. 115–120.

56. Ondrušová B. Measurement Optimization and Analysis of Selected Physical and Chemical Effects on the Ultra-Weak Proton Emission from a Human Hand. CZECH TECHNICAL UNIVERSITY IN PRAGUE, Faculty of Electrical Engineering, Department of Circuit Theory; 2016. Available from: https://dspace.cvut.cz/handle/10467/64862?show=full.

57. Seber G, Wild C. Nonlinear regression. Hoboken, N.J.: Wiley; 2003.

58. Casella G, Berger RL. Statistical Inference. Cengage Learning; 2001. Available from: https://www.amazon.com/Statistical-Inference-George-Casella/dp/0534243126?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0534243126.

59. Anscombe FJ. The Transformation of Poisson, Binomial and Negative-Binomial Data. Biometrika. 1948; 35(3/4):246. https://doi.org/10.1093/biomet/35.3-4.246

60. Kristoufek L. Spectrum-based estimators of the bivariate Hurst exponent. PHYSICAL REVIEW E. 2014; 90(6).

61. Li C. Rescaled-range and power spectrum analyses on well-logging data. GEOPHYSICAL JOURNAL INTERNATIONAL. 2003; 153(1):201–212. https://doi.org/10.1046/j.1365-246X.2003.01893.x

62. Dietrich CR, Newsam GN. Fast and Exact Simulation of Stationary Gaussian Processes through Circulant Embedding of the Covariance Matrix. SIAM J Sci Comput. 1997; 18(4):1088–1107. https://doi.org/10.1137/S1064827592240555

63. Delignières D. Correlation Properties of (Discrete) Fractional Gaussian Noise and Fractional Brownian Motion. Mathematical Problems in Engineering. 2015; 2015:1–7.

64. Weiss M, Elsner M, Kartberg F, Nilsson T. Anomalous Subdiffusion Is a Measure for Cytoplasmic Crowding in Living Cells. Biophysical Journal. 2004; 87(5):3518–3524. https://doi.org/10.1529/biophysj.104.044263 PMID: 15339818

65. Aon MA, Cortassa S. On the fractal nature of cytoplasm. FEBS letters. 1994; 344(1):1–4. https://doi.org/10.1016/0014-5793(94)00321-1 PMID: 8181555

66. Ratto TV, Longo ML. Anomalous Subdiffusion in Heterogeneous Lipid Bilayers†. Langmuir. 2003; 19 (5):1788–1793. https://doi.org/10.1021/la0261803

67. Prasad J, Kopelman R. Fractal-like molecular reaction kinetics: solute photochemistry in porous membranes. Journal of Physical Chemistry. 1987; 91(2):265–266. https://doi.org/10.1021/j100286a007

68. Kopelman R. Fractal reaction kinetics. Science. 1988; 241(4873):1620–1626. https://doi.org/10.1126/science.241.4873.1620 PMID: 17820893

69. Bénichou O, Chevalier C, Klafter J, Meyer B, Voituriez R. Geometry-controlled kinetics. Nature Chemistry. 2010; 2(6):472–477. https://doi.org/10.1038/nchem.622 PMID: 20489716

70. Aon M. CHAOTIC DYNAMICS AND FRACTAL SPACE IN BIOCHEMISTRY: SIMPLICITY UNDERLIES COMPLEXITY. Cell Biology International. 2000; 24(8):581–587. https://doi.org/10.1006/cbir.2000.0572 PMID: 10924231

71. Saxton MJ. Wanted: A Positive Control for Anomalous Subdiffusion. Biophysical Journal. 2012; 103 (12):2411–2422. https://doi.org/10.1016/j.bpj.2012.10.038 PMID: 23260043

72. Martin DS, Forstner MB, Käs JA. Apparent Subdiffusion Inherent to Single Particle Tracking. Biophysical Journal. 2002; 83(4):2109–2117. https://doi.org/10.1016/S0006-3495(02)73971-4 PMID: 12324428

73. Oliver CJ, Pike ER. Measurement of low light flux by photon counting. Journal of Physics D: Applied Physics. 1968; 1(11):1459–1468. https://doi.org/10.1088/0022-3727/1/11/310

74. Foord R, Jones R, Oliver CJ, Pike ER. The Use of Photomultiplier Tubes for Photon Counting. Applied Optics. 1969; 8(10):1975. https://doi.org/10.1364/AO.8.001975 PMID: 20072559

75. Johnson FA, Jones R, McLean TP, Pike ER. Dead-Time Corrections to Photon Counting Distributions. Physical Review Letters. 1966; 16(13):589–592. https://doi.org/10.1103/PhysRevLett.16.589