

A Regularization Approach to Fractional Dimension Estimation

François Roueff, Jacques Lévy Véhel

▶ **To cite this version:**

François Roueff, Jacques Lévy Véhel. A Regularization Approach to Fractional Dimension Estimation. Fractals 98, Oct 1998, Valleta, Malta. inria-00593254

HAL Id: inria-00593254
https://inria.hal.science/inria-00593254

Submitted on 13 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A REGULARIZATION APPROACH TO FRACTIONAL DIMENSION ESTIMATION

F. ROUEFF, J. LÉVY VÉHEL
INRIA Rocquencourt - Projet Fractales
Domaine de Voluceau
B.P. 105 Le Chesnay Cedex, France
Email: francois.roueff@inria.fr, jlv@bora.inria.fr

We propose a new way of evaluating the regularity of a graph of a function f. Our approach is based on measuring the growth rate of the lengths of less and less regularized versions of f. This leads to a new index, that we call regularization dimension, $dim_{\mathcal{R}}$. We derive some analytical properties of $dim_{\mathcal{R}}$ and compare it with other fractional dimensions. A statistical estimator is derived, and numerical experiments are performed, which suggest that $dim_{\mathcal{R}}$ may be computed in a robust way. Finally, we apply the regularization dimension to the study of Ethernet traffic.

Keywords

Fractal dimension, regularization, estimation, multifractals, wavelets.

1 Introduction

There are two main ways of measuring the regularity of a non-differentiable function $f : K \to \mathbb{R}$ where K is a compact set of $\mathbb{R}$. The first one is based on the investigation of the Hölder properties of f[1]. These can be considered from a global point of view, i.e. one seeks the largest $\alpha_g > 0$ such that $\exists C > 0, \forall x, y \in K, |f(x) - f(y)| < C|x-y|^{\alpha_g}$. A local approach is to look only in a neighborhood of x_0 for the largest exponent such that : $\sup_{x,y \in B(x_0,\rho)} \dfrac{|f(x) - f(y)|}{|x-y|^{\alpha_l}}$ is finite and then to make ρ tend to 0. Finally, the pointwise exponent is the largest α_p such that $\sup_{x,y \in B(x_0,\rho)} \dfrac{|f(x) - f(y)|}{\rho^{\alpha_p}}$ is finite.

The second way of evaluating the regularity of f is to measure the dimension of its graph Γ. Again, several definitions exist, and we only mention here the most frequently used ones, in the case of a subset E of $\mathbb{R}^d$[2] :

Definition 1 (Hausdorff dimension) *Let $\{E_i\}_{1 \le i \le \infty}$ be a δ-*cover *of E, i.e. a countable family of sets such that*

$$E \subset \bigcup_{i=0}^{\infty} E_i,\ |E_i| \le \delta,\ E_i \subset \mathbb{R}^d,\ \forall\, i$$

Consider

$$\mathcal{H}^s_\delta(E) = \inf \sum_{i=0}^{\infty} |E_i|^s$$

Let

$$\mathcal{H}^s(E) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(E) = \sup_{\delta > 0} \mathcal{H}^s_\delta(E)$$

$\mathcal{H}^s(E)$ *is called the* s-dimensional Hausdorff outer *measure of* E. *The Hausdorff dimension of* E *is:*

$$\dim_H E = \inf\{s : \mathcal{H}^s(E) = 0\} = \sup\{s : \mathcal{H}^s(E) = \infty\}$$

Finally, one often considers the box dimension which is easier to compute. We assume now that E is bounded:

Definition 2 (Box-counting dimension) *Let* $N_\delta(E)$ *be the smallest number of sets of diameter equal to* δ *which can cover* E. *The* lower *and* upper box-counting *dimension of* E *are respectively defined by*

$$\overline{\dim}_B E = \overline{\lim_{\delta \to 0}} \frac{\log N_\delta(E)}{-\log \delta}$$

$$\underline{\dim}_B E = \underline{\lim_{\delta \to 0}} \frac{\log N_\delta(E)}{-\log \delta}$$

From now on, we will write dim_B in place of $\overline{\dim}_B$.

The following relations always hold:

$$dim_H \Gamma \leq dim_p \Gamma \leq dim_B \Gamma.$$

For many "nice" curves, all these definitions give the same result, which is moreover related to the Hölder exponent. For example, if X is a fractional Brownian motion of index H, then almost surely: $dim_H X = dim_p X = dim_B X = 2 - H$ and $\forall t, \alpha_l(t) = \alpha_p(t) = \alpha_g = H$. For a Weierstrass function $W(t) = \sum_1^\infty \lambda^{-nh} \sin(\lambda^n t), 0 < h < 1$, one has $dim_H W \leq dim_p W = dim_B W = 2 - h$ and $\forall t, \forall t, \alpha_l(t) = \alpha_p(t) = \alpha_g = h$. More generally, a relation of the kind: $dim \Gamma = 2 - h$, where h is a kind of "self-fimilarity" index, holds for many curves, such as some non Gaussian stationary processes [3, 4].

As can be seen from the previous definitions, all the dimensions are somehow related to the geometric properties of Γ, and in particular its length. More precisely, one can think of dim_B as measuring the growth rate of the lengths of approximations of Γ at finer and finer resolutions. We propose in this paper to investigate more thoroughly this idea, and to define a new dimension based on measuring the lengths of regularized versions of Γ. The main advantages of this approach are that it allows us to make use of a whole body of tools developed in regularization theory, to establish insightful links with wavelet theory and multifractal analysis, and finally to obtain robust estimators, the statistical properties of which can be well described.

Section 2 defines the regularization dimension and gives its basic properties. Comparisons with other dimensions and links with wavelet analysis and multifractal theory are developed in section 3. Section 4 describes an estimator of $dim_{\mathcal{R}}$ and shows numerical experiments. Finally, section 5 presents an application to the analysis of Ethernet traffic.

2 Definition and Basic Properties

2.1 Regularization Dimension

Let Γ be the graph of a bounded function $f : \mathbb{R} \to \mathbb{R}$ whose support K is a closed bounded ball. Let $\chi(t)$ be a kernel function of Schwartz class $\mathcal{S}$ such that:

$$\int \chi = 1. \tag{1}$$

Let $\chi_a(t) = \frac{1}{a}\chi(\frac{t}{a})$ be the dilated version of χ at scale a. Let f_a be the convolution of f with χ_a:

$$f_a = f \star \chi_a.$$

Condition (1) insures that χ_a tend to the Dirac distribution and f_a to f in the sense of distributions as a goes to 0. Since $f_a \in \mathcal{S}$, the length of its graph Γ_a on K is finite and given by:

$$\mathcal{L}_a = \int_K \sqrt{1 + f_a'(t)^2}dt \tag{2}$$

Definition 3 (Regularization Dimension) *Define:*

$$dim_{\mathcal{R}}(\Gamma) = 1 + \overline{\lim_{a \to 0}} \frac{\log(\mathcal{L}_a)}{-\log(a)} \tag{3}$$

$dim_R(\Gamma)$ is called the regularization dimension of Γ.

We will see in section 3 that (3) does not depend on the choice of kernel χ and that χ may in fact be chosen among a wider class of functions.

2.2 Basic properties

We now list some basic properties of $dim_{\mathcal{R}}$ which are easy to prove. Recall that f is assumed to be a bounded function on a compact set.

Proposition 1

$$1 \leq dim_R(\Gamma) \leq 2 \tag{4}$$

Proposition 2

$$\sup_a \mathcal{L}_a < \infty \Rightarrow dim_R(\Gamma) = 1$$

We say in that case that Γ is of finite curve length.

Proposition 3 *Let Γ be the graph of $f_1 + f_2$ and $\Gamma_i, i = 1, 2$ the ones of f_i then:*

- $dim_{\mathcal{R}}(\Gamma_1) = dim_{\mathcal{R}}(\Gamma_2) \Rightarrow dim_{\mathcal{R}}(\Gamma) \leq dim_{\mathcal{R}}(\Gamma_1)$.
- $dim_{\mathcal{R}}(\Gamma_1) > dim_{\mathcal{R}}(\Gamma_2) \Rightarrow dim_{\mathcal{R}}(\Gamma) = dim_{\mathcal{R}}(\Gamma_1)$.

If f_1 and f_2 have disjoint supports then:

$$dim_{\mathcal{R}}(\Gamma = \Gamma_1 \cup \Gamma_2) = \max(dim_{\mathcal{R}}(\Gamma_1), dim_{\mathcal{R}}(\Gamma_2)).$$

Proposition 4 *Let l be a Lipschitzian function over K, i.e. :*

$$\exists C, \forall x, y \in K, |l(x) - l(y)| \leq C|x - y|.$$

Let Γ be the graph of f and Γ' the graph of $g(t) = f(t).l(t)$. Then :

$$dim_{\mathcal{R}}(\Gamma') \leq dim_{\mathcal{R}}(\Gamma)$$

If moreover l does not vanish on K then $\frac{1}{l}$ is also Lipschtzian and :

$$dim_{\mathcal{R}}(\Gamma') = dim_{\mathcal{R}}(\Gamma)$$

Proposition 5 *Let Γ be the graph of f, $\Gamma_i, i = 1, 2$ the ones of f_i. If f_1 and f_2 are bounded and compactly supported and $f = f_1 \star f_2$ then f is bounded and compactly supported and :*

$$dim_{\mathcal{R}}(\Gamma) \leq \min(dim_{\mathcal{R}}(\Gamma_1), dim_{\mathcal{R}}(\Gamma_2))$$

If $f_1 \in L^1$ and f_2 is bounded and compactly supported and $f = f_1 \star f_2$, then f is bounded and for any closed bounded ball B :

$$dim_{\mathcal{R}}(\Gamma|_B) \leq dim_{\mathcal{R}}(\Gamma_2)$$

Proposition 6 *Suppose f is globally Hölderian with exponent $H \in]0, 1]$ over K. Then :*

$$dim_{\mathcal{R}}(\Gamma) \leq 2 - H \tag{5}$$

Proposition 7 *Define :*

$$\tilde{\mathcal{L}}_a = \int_K |f'_a(t)| \tag{6}$$

and :

$$\overline{\mathcal{L}}_a = \int |f'_a(t)| = ||f'_a||_{L^1}. \tag{7}$$

Then :

$$\sup_a \mathcal{L}_a < \infty \Longleftrightarrow \sup_a \tilde{\mathcal{L}}_a < \infty \Longleftrightarrow \sup_a \overline{\mathcal{L}}_a < \infty$$

Moreover, if Γ is of infinite curve length then :

$$dim_{\mathcal{R}}\Gamma = 1 + \overline{\lim_{a \to 0}} \frac{\log(\tilde{\mathcal{L}_a})}{-\log(a)} = 1 + \overline{\lim_{a \to 0}} \frac{\log(\overline{\mathcal{L}_a})}{-\log(a)}$$

2.3 Towards a local point of view

Proposition 8 *Let $g : \mathbf{R} \to \mathbf{R}$ be such that :*

1. $g|_K = f$
2. $\exists P$, *polynomial such that* $\forall t \in \overline{K}, |g(t)| < |P(t)|$.

Then :

$$dim_{\mathcal{R}}(\Gamma) = 1 + \overline{\lim_{a \to 0}} \frac{\log(\mathcal{M}_a)}{-\log(a)} = 1 + \overline{\lim_{a \to 0}} \frac{\log(\tilde{\mathcal{M}}_a)}{-\log(a)}$$

where $\mathcal{M}_a = \int_K \sqrt{1 + g'_a(t)^2} dt$ and $\tilde{\mathcal{M}}_a = \int_K |g'_a(t)|$.

In view of the previous proposition, it becomes natural to define a local dimension:

Definition 4 (Local regularization dimension) *Let Γ be the graph of a real function f such that f is locally bounded and bounded by a polynomial over $\mathbb{R}$. Let B be a bounded closed ball and*

$$\mathcal{L}_a(B) = \int_B \sqrt{1 + (f \star \chi_a)'(t)^2} dt \tag{8}$$

Then we define:

$$dim_R(\Gamma|_B) = 1 + \overline{\lim_{a \to 0}} \frac{\log(\mathcal{L}_a(B))}{-\log(a)} \tag{9}$$

Global and local dimensions

From propositions 7 and 8, it follows that if f is of infinite length over B we can as well define:

$$\mathcal{L}_a(B) = \int_B |(f \star \chi_a)'(t)| dt$$

or

$$\mathcal{L}_a(B) = \int_B |((\mathbb{1}_B.f) \star \chi_a)'(t)| dt$$

or

$$\mathcal{L}_a(B) = \int_{\mathbb{R}} |((\mathbb{1}_B.f) \star \chi_a)'(t)| dt = ||((\mathbb{1}_B.f) \star \chi_a)'||_{L^1},$$

where $\mathbb{1}_B$ is characteristic function of B, without changing (9).
As a consequence, *this local dimension can be seen as the global one applied on $\mathbb{1}_B.f$. Thus, any result in the global case still holds in the local case.*

Pointwise dimension

With the same conditions on f as in definition 4, one can then define a pointwise dimension:

Definition 5 (Pointwise Regularization dimension) *Let $M = (x, f(x))$ be in Γ and $B(x, \epsilon)$ be the ball centered on x of radius ϵ.Define:*

$$dim_R(M) = \lim_{\epsilon \to 0} dim_R(\Gamma|_{B(x,\epsilon)}) \tag{10}$$

Note that proposition 3 implies that, $dim_R(\Gamma|_{B(x,\epsilon)})$ is an increasing function of ϵ. Thus, the limit above is always defined.

3 Links between $dim_{\mathcal{R}}$, other fractal indexes and wavelets

3.1 Link with box-counting dimension

Proposition 9 *If f is continuous, the following relation holds:*

$$dim_{\mathcal{R}}\Gamma \leq dim_B\Gamma.$$

Proof The proof makes use of the results of 12.4 in [5] relating dim_B with the oscillations of f. More precisely:

$$\dim_B(\Gamma) = \overline{\lim_{\tau \to 0}}(2 - \frac{\log Var_\tau(f)}{log\tau}), \tag{11}$$

where $Var_\tau(f) = \int_K \mathrm{osc}_\tau(t)dt$ and $\forall t \in K, \tau > 0, \mathrm{osc}_\tau(t) = \sup_{x,y \in B(t,\tau)} |f(x) - f(y)|$.

Now, $\tilde{\mathcal{L}}_a$ can easily be related to the oscillations of f:

$$\forall t \in K, \forall a > 0, f'_a(t) = \int f(x)(\chi_a)'(t-x)dx = \int (f(x) - f(t))(\chi_a)'(t-x)dx$$

To simplify the proof, we suppose that the support of χ is compact and included in $[-1, 1]$. We have:

$$\begin{aligned} |f'_a(t)| &\leq \int_{B(t,a)} |f(x) - f(t)|.|(\chi_a)'(t-x)|dx \\ &\leq \mathrm{osc}_a(t) \int_{B(t,a)} |(\chi_a)'(t-x)|dx \\ &\leq ||\chi'||_{L^1} \frac{\mathrm{osc}_a(t)}{a} \end{aligned} \tag{12}$$

It follows that:

$$\tilde{\mathcal{L}}_a \leq ||\chi'||_{L^1} \frac{Var_a(f)}{a}$$

If $\tilde{\mathcal{L}}_a$ is finite, $dim_{\mathcal{R}}\Gamma$ is 1 and the result holds. Otherwise, $dim_{\mathcal{R}}\Gamma = 1 + \overline{\lim_{a \to 0}} \frac{\log(\tilde{\mathcal{L}}_a)}{-\log(a)}$ and (11) gives the result. □

3.2 Wavelet and regularity

Formula 7 in proposition 7 gives us a direct link with wavelet analysis. When $\mathcal{L}_a$ is infinite, the regularization dimension is directly related to the variations of L^1-norm of the wavelet transform of f as a time function with respect to scale.
Indeed, $f'_a(b)$ is the wavelet transform of f at scale a and time b with $\psi(t) = \chi'(-t)$ as analyzing wavelet. Since we took $\chi \in \mathcal{S}$, ψ has one vanishing moment and is fast decreasing. It also verifies the admissibility condition. One can then make use of classical properties to investigate $dim_{\mathcal{R}}$ from a new point of view. In particular, in a recent study [6], the link between the wavelet coefficients of f and $dim_B\Gamma$ has been established. In the same paper, an example is given of a function f, such that $dim_{\mathcal{R}}\Gamma < dim_B\Gamma$. However, it seems that in "many" cases the two dimensions coincide.

3.3 Link with multifractal analysis

The following function is classical in multifractal analysis [7]:

$$\tau(q) = \underline{\lim_{a \to 0}} \frac{\log \int |C(a,b)|^q db}{\log a},$$

where $C(a,b)$ is the wavelet coefficient of f at scale a and position b using a wavelet of sufficient regularity. Thus, as seen previously, $\int |C(a,b)|db$ is directly related to $\overline{\mathcal{L}}_a$. More precisely, we have obviously:

Proposition 10

$$dim_{\mathcal{R}}\Gamma = \max(1, 2 - \tau(1))$$

Several conclusions can be drawn from this relation. First, as announced in section 2.1, $dim_{\mathcal{R}}\Gamma$ does not depend of the choice of the kernel χ as long as χ' is a wavelet with sufficient regularity (which is always the case for $\chi \in \mathcal{S}$). Second, since it is always true that [1]:

$$\tau(1) = \inf_{\alpha}(\alpha - f_g(\alpha))$$

where f_g is the large deviation multifractal spectrum of f, we set that $dim_{\mathcal{R}}$ is related to multifractal spectrum of f. Third, there are a number of cases where it has been shown that $dim_b\Gamma = 2 - \tau(1)$, e.g. in the case of attractors of affine IFS. In all these situations, we have:$dim_B\Gamma = dim_{\mathcal{R}}\Gamma$.

4 Estimation of Regularization dimension

4.1 Statistical Properties

For practical purposes, it is important to investigate the behavior of $\mathcal{L}_a$ and $dim_{\mathcal{R}}$ when noise is added to the data. In this section, we consider the simple case where the deterministic signal f is corrupted with additive white Gaussian noise b of mean 0 and variance σ^2. To simplify, we will use a discrete version $\mathcal{L}_n$ of $\mathcal{L}_a$ corresponding to the discrete wavelet transform. We will denote $a(n)$ the scale corresponding to the index n. Let $X = f + b$ be the signal to analyzed, $x_n^k = f_n^k + b_n^k$ its wavelet coefficients using an orthonormal wavelet basis, n being the scale index and k being the position index. Let Γ be the graph of f, $\tilde{\Gamma}$ the graph of X, and Γ_b the "graph"[a] of b. Then $\mathcal{L}_n = \sum |c_n^k|$, up to a scale normalization.

It is well known that: $b_n^k \stackrel{d}{=} N(0,\sigma)$, where $\stackrel{d}{=}$ means equal in distribution and $N(0,\sigma)$ denotes a normal law of mean 0 and variance σ^2. Since f is deterministic, $x_n^k \stackrel{d}{=} N(f_n^k,\sigma)$. Note at this point that it is straightforward to show that $\dim_{\mathcal{R}}\Gamma_b = 2.5$ almost surely (a.s.) (For the graph B of a Brownian motion, $dim_{\mathcal{R}}B = 1.5\, a.s.$, and if Z is the graph of the increments of Y, i.e. the derivative distribution, $dim_{\mathcal{R}}Z = dim_{\mathcal{R}}Y + 1$ a.s.).This implies that $dim_{\mathcal{R}}\tilde{\Gamma} = 2.5$ a.s.

In order to access to $dim_{\mathcal{R}}\Gamma$, we thus need to refine our analysis. From an intuitive point of view, the situation is clear: at low resolutions, i.e. for large scales, $\mathcal{L}_n(X) \sim \mathcal{L}_n(f)$, while at fine scales, the noise is predominant and $\mathcal{L}_n(X) \sim \mathcal{L}_n(b)$. A way to obtain a useful estimation is then to evaluate $dim_{\mathcal{R}}$ not as a limit at fine scales but over a restricted ranges of scales, i.e. for $a \leq a(n^\star)$, $a(n^\star)$ depending on

[a] b can be defined as the derivative distribution of a Brownian motion $(B_t, t \in \mathbf{R}_+)$. Thus, it is not a bounded function, and the term of "graph" is not rigorous. Nevertheless, the definition of $dim_{\mathcal{R}}$ still holds although not all the results on bounded functions remain valid. For instance, (4) is not always true for distributions.

the Signal to Noise Ratio (SNR). In the following, we describe a method to compute $n^\star$ and to estimate $\mathcal{L}_n(f)$. An easy calculation leads to:

$$E|x_n^k| = \sigma\sqrt{\frac{2}{\pi}}\exp(-\frac{{f_n^k}^2}{2\sigma^2}) + |f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma})$$

, where E denotes expectation. The first term of the sum above corresponds to the contribution of noise, i.e. $\sigma\sqrt{\frac{2}{\pi}}$, weighted by $\exp(-\frac{{f_n^k}^2}{2\sigma^2})$, which tends to 0 when the SNR goes to infinity, while the second term corresponds to the contribution of f, i.e. $|f_n^k|$, weighted by $erf(\frac{|f_n^k|}{\sqrt{2}\sigma})$, which tends to 0 when the SNR goes to 0. Summing this over all the positions k leads to the same structure for $\mathcal{L}_n(X)$ i.e. $\mathcal{L}_n(X) = A(n) + B(n)$ where $A(n) = \sum_k \sigma\sqrt{\frac{2}{\pi}}\exp(-\frac{{f_n^k}^2}{2\sigma^2})$ and $B(n) = \sum_k |f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma})$.
Now, simple algebra yields: $E\exp(-\frac{{x_n^k}^2}{2\sigma^2}) = \frac{1}{\sqrt{2}}\exp(-\frac{{f_n^k}^2}{4\sigma^2})$. Hence:

$$E|x_n^k| = E\frac{2\sigma}{\sqrt{\pi}}\exp(-\frac{{x_n^k}^2}{\sigma^2}) + |f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma}) \tag{13}$$

Thus, $\tilde{A}(n) = \sum_k \frac{2\sigma}{\sqrt{\pi}}\exp(-\frac{{x_n^k}^2}{\sigma^2})$ is an unbiased estimator of $A(n)$. The ratio $R(n) = \frac{\tilde{A}(n)}{\mathcal{L}_n(X)}$ will give the relative importance of the noise term in $\mathcal{L}_n(X)$. Moreover, $\mathcal{L}_n(X) - \tilde{A}(n)$ is then an unbiased estimator of $B(n)$ which is a lower bound of $\mathcal{L}_n(f)$. Practically, $R(n)$ allows us to find $n^\star$ and formula (13) will let us estimate f_n^k from the x_n^k: the method is now either to estimate $E\{c_n^k = |x_n^k| - \frac{2\sigma}{\sqrt{\pi}}\exp(-\frac{{x_n^k}^2}{\sigma^2})\}$ (problem 1.1) and to find each f_n^k by solving $|f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma}) = Ec_n^k$ (problem 1.2) or to estimate $E\{\sum_k c_n^k\}$ (problem 2.1) and to find $\sum f_n^k$ using $\sum_k |f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma}) = E\{\sum_k c_n^k\}$ (problem 2.2). Among these four problems, two are easy to solve: the problems 1.2 and 2.1 (because $\sum_k c_n^k$ is actually a good estimator of $E\{\sum_k c_n^k\}$).
An approximated solution to problem 1.1 is to estimate Ec_n^k by the mean of the continuous wavelet transform coefficients of X over an interval of the order of magnitude of the scale around the position k. Experiments (see section 4.2) show that this method gives good result. This first method will be referred to the *method of local mean*.
Problem 2.2 may be approximatively solved by assuming that:

$$\sum_k |f_n^k| erf(\frac{|f_n^k|}{\sqrt{2}\sigma}) \sim \sum_k |f_n^k| erf(\frac{\overline{|f_n|}}{\sqrt{2}\sigma}),$$

where $\overline{|f_n|}$ is the mean value of $|f_n^k|$ and is directly related to $\sum_k |f_n^k|$. Experiments show that this assumption is too optimistic. This second method will be referred to as the *method of global mean*.

4.2 Numerical Results

Noise free signal

We estimated $dim_{\mathcal{R}}$ with a Gaussian kernel on 2^{11} points samples of fractional Brownian motions (fBm) and deterministic Weierstrass functions (WF) with varying exponent $H = 0..1$. For an fBm or a WF of constant Holder exponent H, the theoritical regularization dimension of the graph is $2 - H$.

Table 1: This table contains estimated regularization dimensions of fBms and WFs for different Hölder exponents. The scale regression range is the same for all values of H.

H	*Estimated* $dim_{\mathcal{R}}$		$2 - H$
	WF	fBm	
0.1	1.91	1.93	1.9
0.2	1.81	1.77	1.8
0.3	1.71	1.74	1.7
0.4	1.61	1.55	1.6
0.5	1.52	1.49	1.5
0.6	1.43	1.41	1.4
0.7	1.35	1.35	1.3
0.8	1.27	1.26	1.2
0.9	1.20	1.12	1.1

Noisy signal

We applied the different methods explained in section 4.1 with a Gaussian kernel on a Weierstrass function on $[0, 1]$ of Hölder exponent 0.5 corrupted by a white Gaussian noise of $SNR = 10 \log_{10} \frac{\int f^2}{\sigma^2}$ equal to -6.0db (see figure 1).

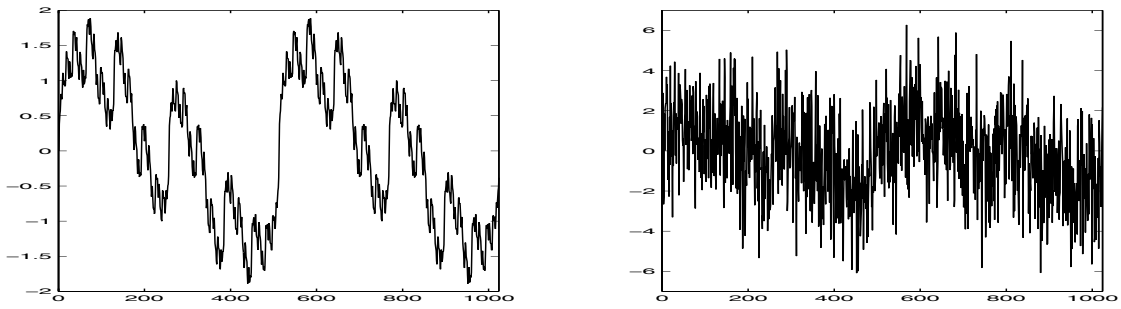


Figure 1: Left: Graph of the Weierstrass function of Hölder exponent 0.5. Right: Graph of the same signal corrupted by a Gaussian noise with $SNR = -6.0db$.

Results are displayed in figure 2: one observes that the different components of $\mathcal{L}_n(X)$ separate from each other when $R(n)$ becomes larger than 10%. When $R(n)$ reaches a plateau close to 1, the estimation of $\mathcal{L}_n(f)$ by the *method of local mean* still gives good results. Hence the *experimental* conclusions:

- For n such that $R(n) < 10\%$, any of the estimations of $\mathcal{L}_n(f)$ gives good results, which is explained by the theoretical analysis of section 4.1.
- For n such that $R(n)$ has not reached a plateau, the *method of local mean* gives the best results and allows to estimate $dim_{\mathcal{R}}$ with good accuracy.

We then have both a method to choose the regression range and to estimate $dim_{\mathcal{R}}$ on data corrupted with additive Gaussian noise. These conclusions have been confirmed by other experiments with different signals and different signal-noise ratios.

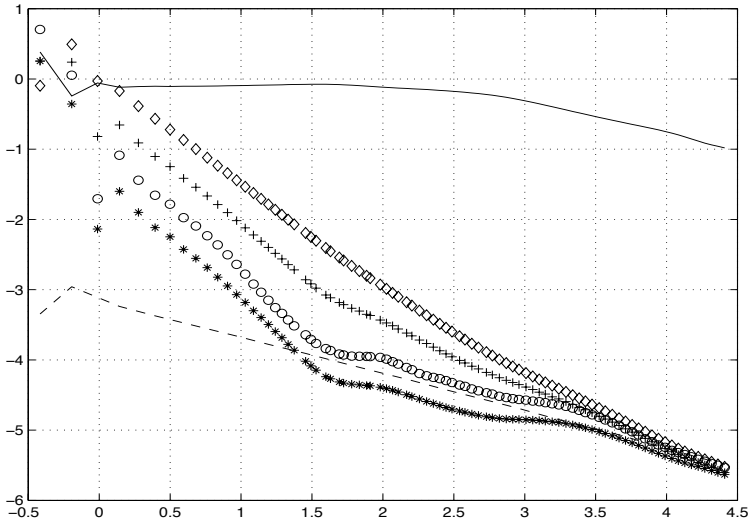


Figure 2: loglog plot of estimators of $\mathcal{L}_n(f)$ versus scale. $\diamond$ are the $\mathcal{L}_n(X)$, the dashed line is $\mathcal{L}_n(f)$ (reference signal), $+$ are the estimations of $\mathcal{L}_n(f)$ by the *method of global mean*, o are the estimations of $\mathcal{L}_n(f)$ by the *method of local mean*, $\star$ are the estimations of $B(n)$. At last, the plain line is the decimal logarithm of $R(n)$

5 Application to the study of Ethernet Traffic

5.1 Introduction

The self-similar nature of Ethernet traffic has been investigated in two different ways: first it has been proven that it has a fractal-like behavior [8]. A different approach [1] showed that Ethernet Traffic has multifractal properties. A model based on superposition of many ON/OFF sources has also been proposed [9]. Here, we apply the regularization dimension on such data, and we investigate the validity of this model.

5.2 Traffic model

Using the notations and the results of the authors of the model [9], let $W^{(m)}(t), t \geq 0$ be the binary time series generated by one source: $W^{(m)}(t) = 1$ means there is a packet at time t, $W^{(m)}(t) = 0$ means there is no packet. For M independent and identically distributed sources, the traffic data is a superposition of all sources packets emission: $W_M(t) = \sum_{m=1}^{M} W^{(m)}(t)$. Define the aggregated cumulative packet traffic as the following time series: $W_M^*(Tt) = \int_0^{Tt} W_M(u)du$, where T is a scale parameter. The model rests on some assumptions on the ON- and OFF periods distributions of each source described by some parameters. These parameters which we won't discuss here define one type of source. Now, assume R types of sources and proportions $M^{(r)}/M$ of sources of type $r = 1, \ldots, R$, with $M^{(r)}/M$ not converging to 0, as $M \to \infty$. Then, one can calculate for each type of sources the positive real parameters $\mu_1^{(r)}, \mu_2^{(r)}, \sigma_L^{(r)}, H^{(r)} < 1$ and a slowly varying function at infinity $L^{(r)}$ (e.g. a constant) such that the following result holds:

Theorem 1 *For large $M^{(r)}, r = 1, \ldots, R$ and large T, the aggregated cumulative packet traffic $W_M^*(Tt)$ behaves statistically like:*

$$T(\sum_{r=1}^{R} M^{(r)} \frac{\mu_1^{(r)}}{\mu_1^{(r)} + \mu_2^{(r)}})t + \sum_{r=1}^{R} T^{H^{(r)}} \sqrt{L^{(r)}(T) M^{(r)}} \sigma_L^{(r)} B_{H^{(r)}}(t) \qquad (14)$$

where $B_{H^{(r)}}(t)$ are independent fBms of exponents $H^{(r)}$.

In terms of increments of $W_M^*(Tt)$, $W_M^*(Tt) - W_M^*(Tu) = \int_{Tu}^{Tt} W_M(u)du$, this theorem tells that they approximatively equal a constant plus a weighted sum of independent fBm increments. Our purpose is now to use this model to analyze Traffic data via the regularization dimension.

5.3 Using Regularization dimension for studying the data

We already gave the result of the regularization dimension calculated on an fBm of Hölder exponent H in section 4.2: $2 - H$. More precisely, at any scale, the mean of the absolute value of the wavelet coefficients at scale a behaves like Ca^{H-1}, where C is a constant. In our case, $\tilde{\mathcal{L}}_a$ applied to a path of an fBm is an estimator of this value. Thus, the linear regression over any range of scales of the loglog plot of $\tilde{\mathcal{L}}_a$ VS a let us access to H.

Now, for a finite weighted sum of independent fBms of different Hölder exponents H_i, let $EC(a)$ be the mean of the absolute value of the wavelet coefficients at a scale a. $EC(a)$ is bounded as follows:

$$C_j a^{H_j - 1} \leq EC(a) \leq \sum_i C_i a^{H_i - 1}$$

for any j. Thus, choosing j such that $C_j a^{H_j - 1}$ is predominant in $\sum_i C_i a^{H_i - 1}$ "around" a given scale $\hat{a}$, the linear regression of the loglog plot of $\tilde{\mathcal{L}}_a$ VS a around $\hat{a}$ let us access to H_j. For instance, the term with smallest H_i will prevail at small scales and the term with highest H_i will prevail at high scales. For the terms with

H_i between these two extremal values, the behavior will depend on the weights C_i. In figure 3, we give an example of this method applied to the sum of two independent fBms of exponents 0.4 and 0.8.

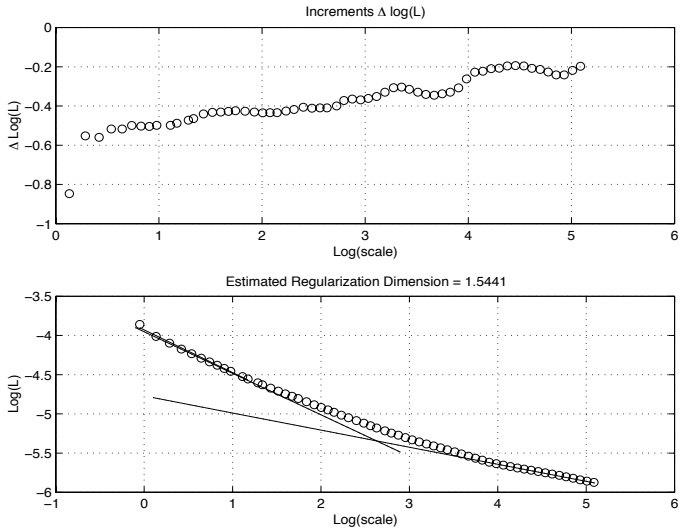


Figure 3: loglog plot of $\tilde{\mathcal{L}}_a$ versus scale and its increments applied to a sum of two independent fBms of exponents $H_{1,2} = 0.4, 0.8$. The chosen range of scales shows the transition between the two slopes corresponding to $H_{1,2}$: $1-0.4 = -0.6$ at small scales and $0.8-1 = -0.2$ at high scales.

We applied this analysis derived from the regularization dimension (we look for regularization dimensions localized in scales) to Ethernet traffic data. We considered a 47813 points sample of Ethernet traffic data measured at Berkeley $W_M(n), n = 1, \ldots, N$. We worked directly on samples of the increments of the series $W_M^*(Tt)$: $\Delta(n) = \sum_{k=T(n-1)+1}^{Tn} W_M(k), n = 1, \ldots, N/T$, which, according to the model, should behave, for T big enough, like a constant plus a sum of increments of independent fBms. The constant is not seen by our method because $\tilde{\mathcal{L}}_a$ uses a convolution of the signal with the derivative of a Schwartz kernel function (here we took a Gaussian kernel). Working on increments just multiplies $\tilde{\mathcal{L}}_a$ by a factor a^{-1} and adds a factor 1 to the regularization dimension. It follows that H equals $3 - dim_{\mathcal{R}}(\Gamma')$, where Γ' is the graph of the increments of an fBm of Hölder exponent H and that $\tilde{\mathcal{L}}_a$ applied to $\Delta(n), n = 1, \ldots, N/T$ should behave like $C_j a^{H_j - 2}$, the predominant term in $\sum_i C_i a^{H_i - 1}$ around a given scale $\hat{a}$. Then, the only preprocessing of the data was to adjust the scale parameter T : on one hand, T has to be big enough for applying the model; on the other hand, the bigger T, the smaller the length of the sample $\Delta(n), n = 1, \ldots, N/T$, which also reduces the range of scales on which the estimator $\tilde{\mathcal{L}}_a$ is reliable. Up to these slight differences, we made the same analysis of the data as for the sum of the two fBms. The result is that the traffic data exhibit a behavior similar to a sum of fBms (compare figure 4 and figure

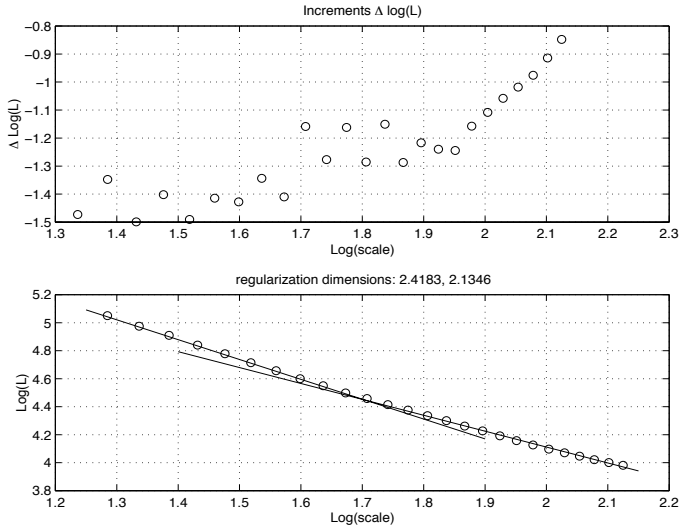


Figure 4: loglog plot of $\tilde{\mathcal{L}}_a$ versus scale and its increments applied to Ethernet traffic data. The chosen range of scales shows the transition between $H_1 = 0.87$ at small scales and $H_2 = 0.6$ at high scales. Hence the two regularization dimensions $\sim 2.4, 2.13$

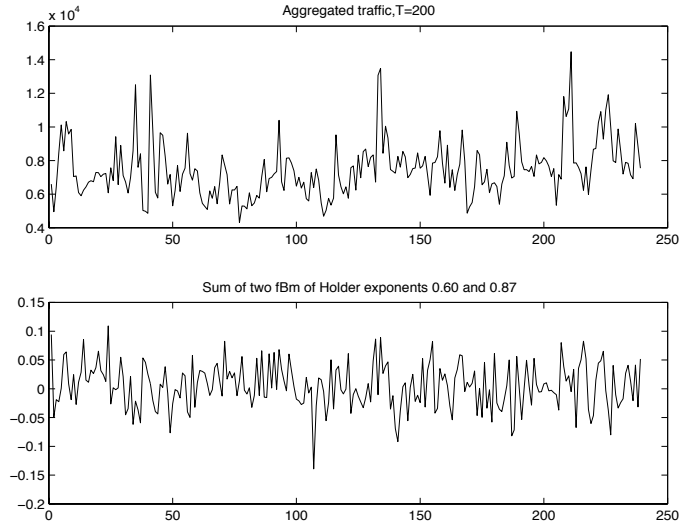


Figure 5: Aggregated traffic and increments of the sum of two fBm of same H_i.

3). Thus, our results do not invalidate the model. Furthermore, they let us find two values $H_{i=1,2} \sim 0.6, 0.87$. We took $T = 200$ which means that Δ had a length $N/T = 239$. To access smaller and higher scales and maybe other $H_{i=3,4,...}$, one should have a longer sample $W_M(n), n = 1, \ldots, N$ and take a higher T.
To give a visual idea of the model we used, we plot the aggregated traffic with $T = 200$ and the increments of a sum of two fBm of Hölder exponents 0.6 and 0.87 on the figure 5.

5.4 Conclusion : Fractal vs multifractal

The multifractal approach takes in account the coexistence of different Hölder exponents along the time axis. Here, because the regularization dimension is a fractal exponent and not a multifractal index, we tracked different Hölder exponents along the scale axis. As the model described in 5.2 seems to be validated by our study, it would be interesting to study the effects of this model in a multifractal approach.

Acknowledgments

We are thankful to Paulo Gonçalves and Bertrand Guiheneuf for enlightening discussions and to Stéphane Jaffard for the release of his preprint paper [6].

References

1. J. Lévy Véhel and R. H. Riedi, Fractional Brownian motion and data traffic modeling: The other end of the spectrum, Fractals in Engineering 97, Eds. J. Lévy Véhel, E. Lutton, C. Tricot, Springer 1997.
2. K.J. Falconer, Fractal Geometry: Mathematical Foundations and Applications, John Wiley and Sons, New York (1990).
3. K. Daoudi, J. Lévy Véhel and Y. Meyer, Construction of continuous functions with prescribed local regularity, to appear in Constructive Approximation.
4. Peter Hall, Rahul Roy, On the relationship between fractal dimension and fractal index for stationary stochastic processes, Annals of Applied Probability 1994.
5. C. Tricot, Curves and Fractal Dimension, Springer-Verlag(1993).
6. S. Jaffard, Sur la dimension de boîte des graphes, preprint.
7. S. Jaffard, Multifractal formalism for functions. I. Results valid for all functions, SIAM J. Math. Anal. 28 (1997), no. 4, 944–970.
8. Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, On the self-similar nature of Ethernet traffic (Extended Version), IEEE/ACM Transactions on networking, Vol. 2,No. 1 February 1994.
9. Walter Willinger, Murad S. Taqqu, Robert Sherman and Daniel V. Wilson, Self-similarity through high-variability : stastistical analysis of Ethernet LAN traffic at the source level (Extended Version), to appear in IEEE/ACM Transactions on networking.