

Bayesian estimation of the self-similarity exponent of the Nile River fluctuation

S. Benmehdi¹, N. Makarava², N. Benhamidouche³, and M. Holschneider²

¹Departement of Mathematics, University of Bourdj-Bouarreridj, Box 64, 34265 Bourdj-Bouarreridj, Algeria

²Institute for Mathematics, University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany

³Departement of Mathematics, University of M'Sila, Box 166, Msila, Algeria

Received: 18 March 2011 – Revised: 22 June 2011 – Accepted: 23 June 2011 – Published: 29 June 2011

Abstract. The aim of this paper is to estimate the Hurst parameter of Fractional Gaussian Noise (FGN) using Bayesian inference. We propose an estimation technique that takes into account the full correlation structure of this process. Instead of using the integrated time series and then applying an estimator for its Hurst exponent, we propose to use the noise signal directly. As an application we analyze the time series of the Nile River, where we find a posterior distribution which is compatible with previous findings. In addition, our technique provides natural error bars for the Hurst exponent.

1 Introduction

Many geophysical systems exhibit non-trivial multi-scale correlation structures. In particular fractional Brownian motion and fractional Gaussian noise are often found to explain quite well the heterogeneity and multi-scale properties of geophysical time series in particular those from hydrology. In addition this kind of model is capable to discriminate between long and short term dependency. This difference has observable consequences since for instance a long memory process might explain the patterns observed in sedimentary deposits in river run-off areas (see e.g. Millen and Beard, 2003). For this reason it is important to devise suitable techniques to estimate the underlying self-similarity exponent, known as the Hurst exponent. Moreover it is important to provide methods to quantify the uncertainty of the so obtained measurements. We choose in this paper a Bayesian approach, which provides in a natural way both, a mean to

estimate the quantity of interest as well as a way to assess the uncertainty of its value.

Fractional Gaussian noise (FGN) is a Gaussian stochastic process $\{G_t^H, t \geq 0\}$, that can formally be viewed as the derivative of a fractional Brownian motion (FBM) $B_t^H, t \geq 0$. This is the only Gaussian process with stationary increments which is self-similar and of zero mean. The parameter H characterizes its behavior under rescaling

$$B^H(at) \simeq a^H B^H(t).$$

Here \simeq denotes equality of distributions. The first and second moments fully characterize this process

$$\begin{aligned} \mathbb{E}(B^H(t)) &= 0, \\ \mathbb{E}(B^H(t)B^H(u)) &= \frac{1}{2}(|t-u|^{2H} - |t|^{2H} - |u|^{2H}). \end{aligned}$$

The trajectories are continuous functions. However often time series of real data represent much noisier features, which however may have hidden self-similar correlation structures with sometimes long-range dependencies. See Fig. 7 for an example showing the level of Nile River from the years 622 to 1284 AD. For this consider the discrete process of increments over fixed time interval $\Delta t = 1$

$$G_i^H = B_i^H - B_{i-1}^H, (i = 1, 2, \dots).$$

The choice of time interval $\Delta t = 1$ is no loss of generality, since any other time interval would lead to the same process up to some multiplicative amplitude factor. This process is the coarse grained fractional Gaussian noise process FGN. It is a stationary process, which is characterized by its auto correlation function. Using the correlation of the FBM we see that



Correspondence to: S. Benmehdi
(sabah_benmehdi@yahoo.fr)

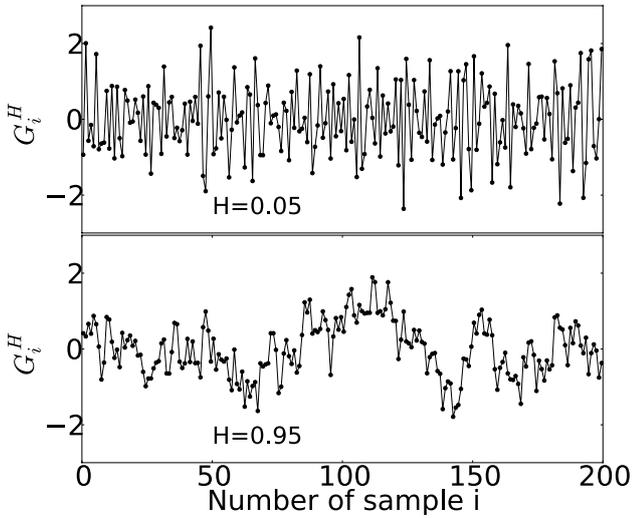


Fig. 1. (Top) Simulation of fractional Gaussian noise for $N = 200$ and $H = 0.05$. (Bottom) Simulation of fractional Gaussian noise for $N = 200$ and $H = 0.95$. The difference in correlation structure is clearly visible.

$$\begin{aligned} \rho(k) &= \mathbb{E}(G^H(i+k)G^H(i)) \\ &= \frac{1}{2}(|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}), \quad k \in \mathbb{Z}. \end{aligned} \tag{1}$$

For $H = 1/2$ we have the independent identically distributed Gaussian variables and uncorrelated time series after one single time step. For $H \neq 1/2$ however, as the time lag k gets large, the correlation function decays asymptotically like

$$\rho(k) \sim H(2H-1)|k|^{2H-2}, \quad k \rightarrow \infty. \tag{2}$$

For this reason, H quantifies the correlation at large times. For $H < 1/2$ correlation is negative, whereas for $H > 1/2$ the correlation is positive. See Fig. 1 for an example of FGN.

Various methods to estimate the self-similarity exponent H from a time series w_n of observations have been proposed. A standard method consist in replacing w_n by the associated random walk

$$S_n = \sum_{k \leq n} (w_k - \bar{w}), \quad \bar{w} = \frac{1}{N} \sum w_k$$

and to apply techniques to estimate the Hurst exponent of fractional Brownian motion. Here the following types of methods are proposed in the literature. The methods can be classified as temporal, spectral and time-scale methods. The temporal methods are e.g. aggregated variance method (Taquu et al., 1995), detrended fluctuation analysis (Goldberger et al., 2000; Peng et al., 1994), and range scaled analysis (Mandelbrot and Wallis, 1969). From the group of spectral methods, we would like to mention the log-periodogram

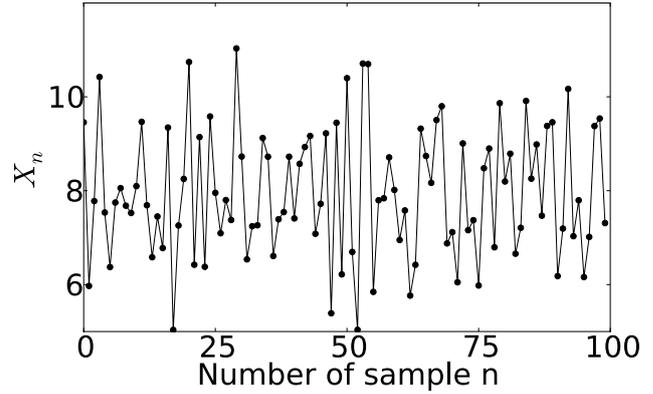


Fig. 2. A random sample of noisy signal $X_n = 1.5G_n^{0.3} + 8.0$, $n = 1, \dots, N, N = 100$.

method (Geweke and Porter-Hudak, 1983), the modified periodogram method, and the Whittle estimator (Whittle, 1953). The third group is in the wavelet domain, which includes the wavelet maximum likelihood WML estimator (McCoy and Walden, 1996) and the Abry-Veitch Daubechies wavelet-based estimator (Abry and Veitch, 1998).

In this paper we want to introduce a Bayesian estimator for the Hurst exponent H of FGN. This is an extension of the method proposed in (Makarava et al., 2011), where we introduced a Bayesian estimator for the Hurst exponent of fractional Brownian motion. One big advantage of our approach compared to others will be, that we take fully into account the correlation structure present in fractional Gaussian noise. Moreover, we will be able to produce error bars for all quantities, that we estimate.

2 Definition of the model

Often the data comes with an unknown offset. Instead of removing it prior to the analysis, we will incorporate it into the model. So we consider discrete time series of the form

$$X_n = \lambda G_n^H + \beta, \quad n = 1, 2, \dots, \tag{3}$$

where G_n^H is the FGN and $H \in]0, 1[$ is the Hurst exponent, $\lambda > 0$ and $\lambda \in \mathbb{R}$ is the amplitude, and $\beta \in \mathbb{R}$ is the offset. Suppose we are given the observations $X_n, n = 1, \dots, N$, the problem we want to address is how to obtain estimators for all involved parameters. For this we propose to use Bayes theorem, that reads

$$\begin{aligned} \mathbb{P}(\beta, \lambda, H | \{X_n\}, n = 1, \dots, N) = \\ C L(\beta, \lambda, H | \{X_n\}_{n=1, \dots, N}) \mathbb{P}(\beta, \lambda, H), \end{aligned}$$

where $L(\beta, \lambda, H | X_n) = \mathbb{P}(\{X_n\}_{n=1, \dots, N} | \beta, \lambda, H)$ is the likelihood function, C is a normalization constant, and $\mathbb{P}(\beta, \lambda, H)$ is some prior information we might have about the parameters. In the absence of such information, we suggest to use

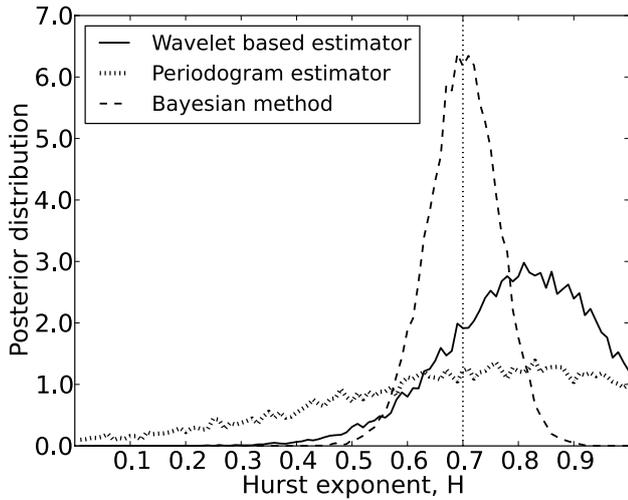


Fig. 3. The distribution of the estimator of the Hurst exponent for $H = 0.7, N = 128$ with 20 000 realizations by wavelet based estimator (solid line), periodogram method (dotted line) and Bayesian method (dashed line).

uninformative priors, $\mathbb{P}(\beta, \lambda, H) \simeq \lambda^{-1}$ where as usual, we use the Jeffreys prior for the scale parameter λ .

For fixed parameters, the observations $\mathbf{X} = [X_1, \dots, X_N]^t$ are multivariate Gaussian random variables with mean value and covariance given by

$$\mathbb{E}(\mathbf{X}) = \mathbf{F}\beta, \quad \mathbb{E}(\mathbf{X}\mathbf{X}^t) = \Sigma,$$

with Σ given by Eq. (1) and $\mathbf{F}^t = [1, \dots, 1]$ is the vector with N components, where $[\cdot]^t$ denotes the transpose. The likelihood function for the parameters λ, β, H can now be written as

$$L(\lambda, \beta, H|\mathbf{X}) = \frac{1}{(2\pi)^{N/2} \lambda^N |\Sigma|^{1/2}} e^{-(\mathbf{X}-\mathbf{F}\beta)^t \Sigma^{-1} (\mathbf{X}-\mathbf{F}\beta)/2\lambda^2}. \quad (4)$$

From Bayes theorem, we now obtain the following posterior density of our parameters

$$\mathbb{P}(\lambda, \beta, H|\mathbf{X}) = C \frac{1}{\lambda^{N+1} |\Sigma|^{1/2}} e^{-(\mathbf{X}-\mathbf{F}\beta)^t \Sigma^{-1} (\mathbf{X}-\mathbf{F}\beta)/2\lambda^2}. \quad (5)$$

The normalization constant C is chosen to make the posterior probability density integrate to one. Then the numerator of the exponent can be written as the following quadratic polynomial in β

$$\mathbf{F}^t \Sigma^{-1} \mathbf{F} \beta^2 - 2\beta \mathbf{F}^t \Sigma^{-1} \mathbf{X} + \mathbf{X}^t \Sigma^{-1} \mathbf{X}$$

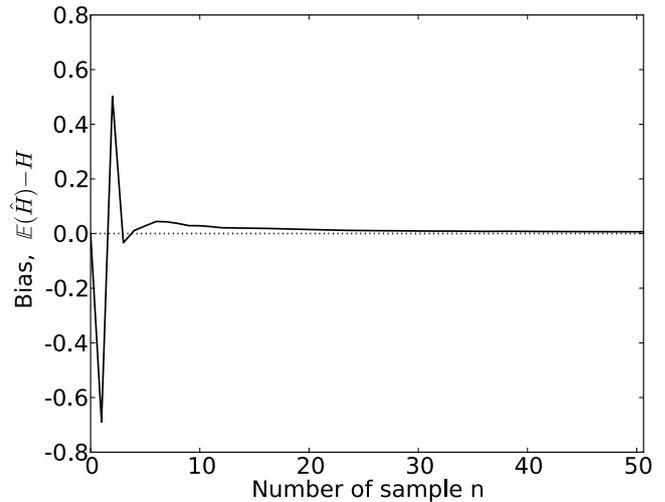


Fig. 4. The validation test for Bayesian estimation with 500 realizations for the Hurst exponent $H = 0.7$.

Table 1. The interval for \hat{H} with $p \geq 90\%$ for synthetic data with $H = 0.7$ and $N = 128$.

Method	\hat{H}	Interval for \hat{H} with $p \geq 90\%$
Wavelet based estimator	0.81	[0.57, 1.05]
Periodogram estimator	0.83	[0.21, 1.31]
Bayesian approach	0.71	[0.59, 0.81]
True value	0.7	

From this we see, that the posterior can be written in the following ‘‘Gaussian’’ form

$$\mathbb{P}(\lambda, \beta, H|\mathbf{X}) = C \frac{1}{\lambda^{N+1} |\Sigma|^{1/2}} e^{-R^2/2\lambda^2} e^{-\frac{\gamma^2(\beta-\beta^*)^2}{2\lambda^2}} \quad (6)$$

$$= C' \frac{e^{-R^2/2\lambda^2}}{\gamma \lambda^N |\Sigma|^{1/2}} f_{\beta^*, \lambda^2/\gamma^2}(\beta). \quad (7)$$

Here f_{μ, σ^2} is the one dimensional Gaussian probability density function with mean μ and variance σ^2 . Note that $\Sigma \Sigma_H$ and therefore, the residuum R^2 and the mode β^* depend on H via

$$\gamma^2 = \mathbf{F}^t \Sigma_H^{-1} \mathbf{F} \quad (8)$$

$$\beta^* = \frac{\mathbf{F}^t \Sigma_H^{-1} \mathbf{X}}{\gamma^2}, \quad (9)$$

$$R^2 = \mathbf{X}^t \Sigma_H^{-1} \mathbf{X} - \gamma^2 \beta^{*2}. \quad (10)$$

Equation (6) describes the full posterior information that we have about all the parameters jointly. In case, we are interested in only one of them, we may treat the other parameters as completely unknown and integrate them out, to obtain the marginal distributions. Integrating over λ we obtain

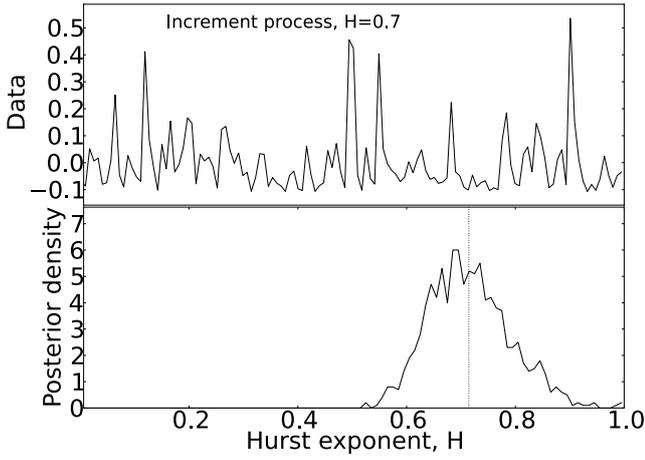


Fig. 5. (Top) Simulation of Rosenblatt process for $N = 129$ and $H = 0.7$. (Bottom) The averaged of the obtained posterior densities for point estimator derived with Bayesian method for 1000 such realizations.

$$\mathbb{P}(\beta, H | \mathbf{X}) = \frac{C_1}{|\boldsymbol{\Sigma}_H|^{1/2} (R^2 + \gamma^2 (\beta - \beta^*)^2)^{N/2}}. \quad (11)$$

The integral over the offset yields

$$\mathbb{P}(\lambda, H | \mathbf{X}) = \frac{C_2}{\gamma(H) \lambda^N |\boldsymbol{\Sigma}_H|^{1/2}} e^{-R(H)^2 / 2\lambda^2}. \quad (12)$$

The integral over H reads

$$\mathbb{P}(\beta, \lambda | \mathbf{X}) = C_3 \lambda^{-N-1} \int_0^1 \frac{e^{-(R(H)^2 + \gamma^2(H)(\beta - \beta^*(H))^2) / 2\lambda^2}}{|\boldsymbol{\Sigma}_H|^{1/2}} dH.$$

Integrating Eq. (12) over λ yields the posterior distribution of H that we can infer from the observations

$$\mathbb{P}(H | \mathbf{X}) = \frac{C_4}{\gamma(H) |\boldsymbol{\Sigma}_H|^{1/2} R^{N-1}}. \quad (13)$$

In practice, this integral has to be performed numerically. Taking the position of the maximum of this posterior density, we obtain the maximum posterior estimate of the scaling exponent H . In the same way we may obtain posterior estimates of λ and of β

$$\mathbb{P}(\lambda | \mathbf{X}) = C_5 \lambda^{-N} \int_0^1 \frac{e^{-R(H)^2 / 2\lambda^2}}{\gamma(H) |\boldsymbol{\Sigma}_H|^{1/2}} dH \quad (14)$$

and

$$\mathbb{P}(\beta | \mathbf{X}) = C_6 \int_0^1 \frac{e^{-\gamma(H)^2 (\beta - \beta^*)^2}}{R(H)^N |\boldsymbol{\Sigma}_H|^{1/2}} dH. \quad (15)$$

From these expressions we also may produce point estimators. For instance we may set

$$\hat{H} = \operatorname{argmax} \mathbb{P}(H | \mathbf{X}). \quad (16)$$

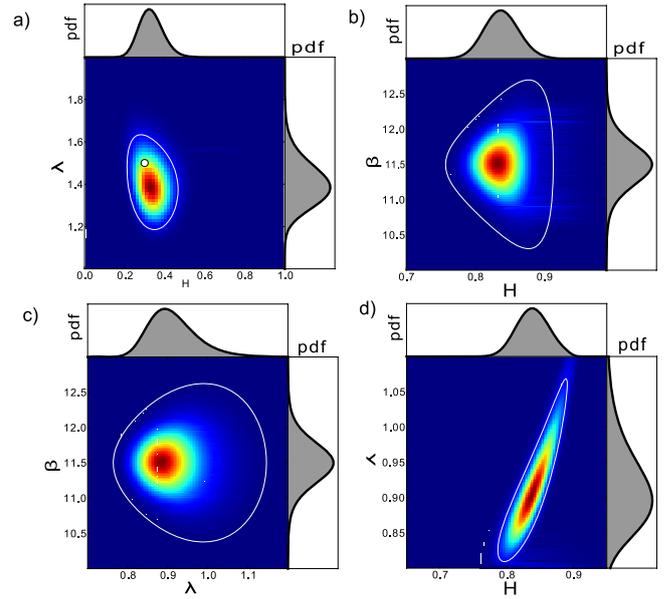


Fig. 6. Normalized two dimensional marginal posterior densities. The maxima indicate the most likely estimates in: (a) $H - \lambda$ plain for signal $X_n = 1.5 G_n^{0.3} + 8.0$, $n = 1, \dots, N = 100$. Note that the posterior is well localized around the point corresponding to the true value (white spot); (b) $H - \beta$ plain for Nile River; (c) $\lambda - \beta$ plain for Nile River; (d) $H - \lambda$ plain for Nile River; on the axis, the one dimensional projections of the posterior densities are depicted. The white contour-line encloses 90 % of posterior probability. It therefore quantifies the posterior uncertainty of the parameters together with their posterior dependency.

This in turn yields an estimator for the offset

$$\hat{\beta} = \beta^*(\hat{H}). \quad (17)$$

3 Application to synthetic data

In this section, we want to show, how to apply the analysis to synthetic data. In the first part, these are data of fractional Gaussian noise. In the second part, we will investigate synthetically generated Rosenblatt processes as examples of non-Gaussian H -self-similar processes.

3.1 Fractional Gaussian noise

For the investigation of fractional Gaussian noise, we generate a random sample

$$X_n = 1.5 G_n^{0.3} + 8.0, \quad n = 1, \dots, N, N = 100. \quad (18)$$

Several methods have been proposed to numerically generate such a realization of FGN, and we refer for more details to (Dieker, 2004; Kang, 2008) and the reference therein.

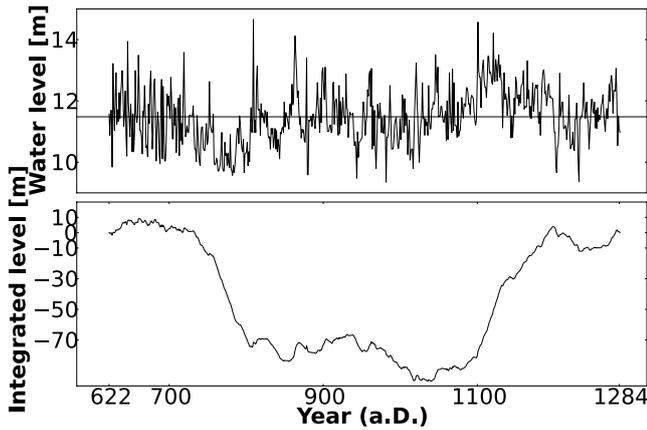


Fig. 7. (Top) The time series of minimal water levels of Nile River near Cairo. (Bottom) The integrated time series.

Table 2. Confidence intervals on the estimated parameters.

Parameter	Estimate	Confidence $\geq 90\%$
\hat{H}	0.83	[0.78, 0.87]
$\hat{\beta}$	11.51	[10.9, 12.1]
$\hat{\lambda}$	0.89	[0.81, 1.04]

Our simulations are based on the use of the Cholesky decomposition of $\Sigma_H = L'L$ where we then apply $L'e$, where $e = [e_1, \dots, e_N]^T$ is a random vector of N independent standard Gaussian random variables. In Fig. 2 we have shown this data.

We then have computed the Bayesian posterior distribution of λ and H together with one of their marginal distributions (see Fig. 6a). As you can see, posterior information is well localized around the data generating parameters. However, since the posterior itself is computed from a random realization its maximum does not coincide with the true value but randomly fluctuates around it.

Next, we use the wavelet based joint estimator by Veitch and Abry (1999) and the periodogram estimator by Robinson (1995) to compare with the point estimator we propose in this work (Eq. 16). We perform 20 000 realizations by Monte-Carlo simulations for the fixed Hurst exponent $H = 0.7$ and with length $N = 128$. For each of them we produce the maximum posterior estimator \hat{H} of our method, wavelet based joint estimator and periodogram method estimator. We implement the wavelet based joint estimator as is proposed by Veitch and Abry (2002). Figure 3 shows the comparison between the methods and in Table 1 the intervals containing $\geq 90\%$ of the distribution are presented. It is clearly depicted that the Bayesian method outperforms the wavelet and periodogram methods.

To make the validation test of the proposed method we quantify the bias of the maximum posterior estimator \hat{H} ,

Table 3. Estimation of the Hurst exponent for different methods.

Estimator	\hat{H}
MPE	0.85
PE	0.90
Whittle	0.84
WML	0.82
FBM	0.80
Bayesian Method	0.83

$\mathbb{E}(\hat{H}) - H$, as a function of the number of data points. For that we generate 500 realizations of fractional Gaussian noise starting from one single observation point. In Fig. 4 it is shown that even starting from 20 data points, the bias quickly decays.

3.2 Rosenblatt process

In this section the Rosenblatt process as an example of a H -self-similar process, whose finite-dimensional distributions are non-Gaussian, is considered. The Rosenblatt process (see Rosenblatt, 1961) is the H -self-similar process with Hurst exponent $H \in (1/2, 1)$ and stationary increments. It can be written in explicit form as a stochastic integral

$$Z(t) = a(H) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\int_0^t (s - y_1)_+^{-\frac{2-H}{2}} (s - y_2)_+^{-\frac{2-H}{2}} ds \right) dB(y_1) dB(y_2),$$

where $B(y), y \in \mathbb{R}$ is pure Brownian motion, $a(H)$ is a positive normalization constant chosen such that $\mathbb{E}(Z(1)^2) = 1$, and $x_+ = \max\{x, 0\}$. We implement the Matlab code from Abry and Pipiras (2006) for the wavelet-based synthesis of the Rosenblatt process and derive the increments of the generated processes. In Fig. 5 we show the example of the increment process for the Rosenblatt process with Hurst exponent $H = 0.7$ of the length $N = 129$. The obtained averaged posterior densities for point estimator derived with Bayesian method for 1000 realizations is also shown. The interval containing $\geq 90\%$ of the distribution here is [0.596, 0.838].

4 Hurst exponent analysis of Nile River

We have applied our technique to the time series of the annual minimum water level of the Nile River for the years 622–1284 AD (663 observations), measured at the Roda Gauge near Cairo. The data set we analyse is publicly available at StatLib archive: <http://lib.stat.cmu.edu/S/beran>. In Fig. 7 we have plotted the time series together with its first discrete integral that is defined by

$$L(k) = \sum_{i=1}^{i=k} (X_i - \bar{X}), \quad k = 1, \dots, N. \quad (19)$$

This data set was studied before by several authors to estimate the Hurst exponent. For example in Liu et al. (2009) the modified Periodogram Estimators (MPE) is used and a comparison with the following methods is presented: the Periodogram Estimator (PE), the Whittle estimator, the wavelet maximum likelihood (WML) estimator and estimators based on the associated FBM.

We have computed the posterior distribution according to Eq. (6). For reasons of visualization, we have computed the marginals over H , λ and β , which are then functions of the remaining two variables only. The results are visible in Fig. 6b–d). We thus find the following posterior estimates (maximum of posterior) of our parameters together with their 90 % confidence intervals (Table 2).

In Table 3 we give a summary of the known results from (Liu et al., 2009), and how they compare to this study. We see that the values obtained by the other methods are in the confidence interval of our result, except for the result obtained by the periodogram estimator (PE). In addition, our method provides estimations of all parameters involved in the model and we also obtain error bars for each of them.

5 Conclusions

In this paper we have proposed a Bayesian estimation technique of the Hurst parameter for the fractional Gaussian noise process. We have considered a slightly more general model, where in addition we have taken the offset and the amplitude as parameters. This technique yields, besides the point estimations of the model's parameters, confidence intervals that enclose a given percentage of the posterior uncertainties. The method was tested successfully on synthetic realizations of fractional Gaussian noise processes. In addition, we performed tests on synthetic realizations of the Rosenblatt process as an example for a non-Gaussian H -self-similar process. Our method turned out to give good results also for Rosenblatt processes. But we can not generalize this to any non-Gaussian self-similar processes. When applied to the historical time series of the Nile River, our results are compatible with previous findings. However in addition we are able to provide error bars for all estimates.

Edited by: J. Kurths

Reviewed by: two anonymous referees

References

- Abry, P. and Pipiras, V.: Wavelet-based synthesis of the Rosenblatt process, *Signal Process.*, 86, 2326–2339, doi:10.1016/j.sigpro.2005.10.021, 2006.
- Abry, P. and Veitch, D.: Wavelet Analysis of Long-Range-Dependent Traffic, *IEEE T. Inform.*, 44, 2–15, 1998.
- Dieker, T.: Simulation of Fractional Brownian motion, Ph.D. thesis, Twente University, 2004.
- Geweke, K. and Porter-Hudak, S.: The estimation and application of long memory time series models, *Time Ser. Anal.*, 4, 221–238, 1983.
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., and Stanley, H.: Components of a New Research Resource for Complex Physiologic Signals, *PhysioBank, PhysioToolkit, and Physionet, Circulation*, 101(23), e215–e220, 2000.
- Kang, D. S.: Simulation of The Fractional Brownian Motion, Ph.D. thesis, J. W. Goethe University, 2008.
- Liu, Y., Liu, Y., Wang, K., Yang, L., and Jiang, T.: Modified periodogram method for estimating the Hurst exponent of fractional Gaussian noise, *Phys. Rev. E*, 80(6), 066207–066207, 2009.
- Makarava, N., Holschneider, M., and Benmehdi, S.: Bayesian estimation of self-similarity exponent, *Phys. Rev. E*, submitted, 2011.
- Mandelbrot, B. B. and Wallis, J. R.: Computer experiments with fractional Gaussian noises. Part 2, rescaled ranges and spectra, *Water Resour. Res.*, 5(1), 242–259, 1969.
- McCoy, E. and Walden, A.: Wavelet Analysis and Synthesis of Stationary Long-Memory Processes, *Comput. Graph. Stat.*, 5, 26–56, 1996.
- Millen, S. and Beard, R.: Estimation of the Hurst Exponent for the Burdekin River using the Hurs-Mandelbrot Rescaled Range Statistic, in: *Proceedings of the First Queensland Statistics Conference*, 2003.
- Peng, C.-K., Buldyrev, S., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A.: Mosaic organization of DNA nucleotides, *Phys. Rev. E*, 49, 1685–1689, 1994.
- Robinson, P.: Log-Periodogram Regression of Time Series with Long Range Dependence, *Ann. Stat.*, 23, 1048–1072, 1995.
- Rosenblatt, M.: Independence and dependence, In *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 2, 431–443, 1961.
- Taqqu, M., Teverovsky, V., and Willinger, W.: Estimators for long range dependence: an empirical study, *Fractals*, 3, 785–798, 1995.
- Veitch, D. and Abry, P.: A Wavelet Based Joint Estimator of the Parameters of Long-Range Dependence, *IEEE Trans. Info. Theory*, special issue “Multiscale Statistical Signal Analysis and its Applications”, 45, 878–897, 1999.
- Veitch, D. and Abry, P.: LDestimate, available at: <http://www.cubinlab.ee.unimelb.edu.au/~darryl/>, last access date: 28. June 2011, 2002.
- Whittle, P.: Estimation and information in stationary time series, *Ark. Mat.*, 2, 423–434, 1953.